WINTERCORP REPORT

BY RICHARD WINTER

SPONSORED RESEARCH PROGRAM

WINTER CORPORATION

The Modern Data Warehouse

Key Concepts and The Cloudera Approach

EXPERTS IN ANALYTIC DATA MANAGEMENT AT SCALE

WINTER CORPORATION

The Modern Data Warehouse

Key Concepts and The Cloudera Approach

RICHARD WINTER

WC /V2

WinterCorp

www.wintercorp.com

42 DERBY LANE, TYNGSBORO, MA

617-695-1800

EXPERTS IN ANALYTIC DATA MANAGEMENT AT SCALE

Summary

THE DATA WAREHOUSE CONCEPT emerged around 1990 and led to widespread implementation of data warehouses and data marts by about 2005.

While data warehouses have met many important business needs for data and decision making, the business world has changed substantially in the course of the last 15 years. Many data warehouse owners are now confronting requirements that these earlier programs were never designed to address. What is needed now is a new, more modern concept of the data warehouse. The modern data warehouse supports data management and analytic requirements that were barely imagined when the original concept took form. Cloudera now offers a data warehouse for the requirements of that companies face today and in the coming years.

On the basis of its independent research, WinterCorp recommends that companies investing in a modern data warehouse look closely at the Cloudera Data Warehouse. Among the large scale data warehouse providers in the market today, Cloudera is unique in offering a complete enterprise data cloud, delivered on prem and on major cloud platforms. The Cloudera Data Platform, of which the Cloudera Data Warehouse is one component, supports all analytic workloads from the Edge to AI, all with enterprise-grade security and governance. ●

Table of Contents

1 Introduction

Few have anticipated the extent to which data and analytics are transforming the modern enterprise today. In every industry, companies are turning themselves into data-driven enterprises. Often these companies have major programs in data science, artificial intelligence or machine learning. Hundreds of large companies have a high-level executive leading a program to transform the company into a "digital" business. Hundreds more have hired a Chief Data Officer, a Chief Analytics Officer or a Chief Digital Officer to signal their serious intent to move ahead in one of these areas. All of these initiatives have one thing in common: as data-driven enterprises, they plan to use data and analytics to implement major changes in the way the companies do business.

As we enter the 2020s, there is a new intensity driven in part by the seemingly sudden emergence of data science and machine learning as practical business tools now being used frequently to predict and prescribe action: factors not present twenty years ago, when "competing on analytics" emerged as a trend. Also not widely present twenty years ago were "big data" and the "Internet of Things (IoT)" that are having such a large impact on modern life, both in the enterprise and at home.

These trends and others mean that an enterprise that wants to compete in the 2020s is going to look very different from the enterprises of twenty years ago. And the enterprise of the 2020s is going to be using data and analytics in a very different way. Companies looking forward into the coming decade, therefore, are now turning attention to a long-standing mainstay in their analytic data program: the data warehouse. The world has changed a lot in the last twenty years, but in many companies, the data warehouse is much the same.

So, at this moment, many are asking: do we need a new approach to the data warehouse for the 2020s? Is it, perhaps, time to create a **modern data warehouse**? One that is suited to the challenges ahead?

This report is about the modern data warehouse: why we need it, what it is, and how to approach creating one. In addition, this report will discuss Cloudera's answer to the need for a modern data warehouse: *a product offering called the Cloudera Data Warehouse, along with WinterCorp's conclusions and recommendations on the larger subject of the modern data warehouse.*

2 Data in the Modern Enterprise

One of the most dramatic ways in which enterprises today look different than they did twenty years ago is in the data they are using for analytic purposes. That data is enormously larger; it is much more varied; and, it is much faster moving than the data in most enterprises in 2000.

A. DATA VARIETY

Throughout the history of data warehousing the data store was a tabular relational database: a collection of tables, where each table consisted of named columns and each row contained the values for those columns. For the most part, the database engines used to run these data warehouses required that each column be limited to values of one of a short list of types, such as integer, float, character, date, currency and so on. The variety and structure of data that could be stored was quite limited. This is still true in the majority of data warehouses running today: some provide for longer text fields and undefined binary objects usually called "blobs," but few really come to grips with the variety of data now present in the digital enterprise.

In the early days of data warehousing, this was seen as perfectly fine: the data used for corporate decision making was principally about sales volumes, prices, orders and so on.

Today, the situation is drastically different. Decision making in businesses is often based on the sentiment expressed in emails, social media posts, calls to the call center, blogs and shopping behavior in stores. Thus free-form text is an important type of data; and, so is audio; so is video. But the variety goes beyond the form of the data to its structure. Therefore, in assessing the significance of a social media post, it may be necessary to understand the network of people who are influenced by the individual making the post: this is the analysis of network data. The structure of network data can be represented in relational tables, but this is not an efficient way to represent it for purposes of analysis; rather, genuine network structures are required.

The same can be said of bill of materials data, which consists of a deep hierarchy; and, many other forms of scientific and engineering data, which have their own structure.

As a result , today's business seeks to make decisions based on data that is extremely varied in type and structure — a situation drastically different from what we saw in practice, when data warehousing was first widely adopted.

B. NEW DATA SOURCES (INCLUDES IOT)

Just as data variety has grown dramatically, the number of new data sources relevant to a business has increased at an even greater pace — and that pace is continuing, probably accelerating.

Whereas the early data warehouse typically drew data from one or two dozen operational systems within the company, modern systems are now often receiving data flows from thousands or tens or thousands of sources. Take sensor data: it is reported that a modern commercial aircraft has as many as ten thousand sensors, each emitting data and looking very much like a separate source to the data warehouse.

In fact, the entire phenomenon of widespread sensor data — now often referred to as the "Internet of Things," or IoT — is new since the emergence of the data warehouse. During the 2020s it will be a rare company that doesn't have a substantial, continuously growing array of sensors feeding data into its data warehouse. Sensors, imaging devices, mobile devices, wearable devices — and other varieties of emitters we can't necessarily imagine today — will all be feeding data into the data warehouse, to help the enterprise understand and manage every aspect of its business operation.

A company's websites and web applications are another major source of data that was not originally seen as part of the data warehouse and is only occasionally treated that way today. But as external and internal websites and web applications have become a pervasive aspect of operating an enterprise, it turns out that their logs can reveal important insights into the behavior, interests, capabilities and needs of customers, partners, suppliers and employees.

The same can be said of social media.

Taken together, there are many more sources of data relevant to corporate decision making than there were in the early stages of data warehousing. This is a trend likely to continue and accelerate in the coming decade.

C. CONTINUOUS DATA UPDATE

The early data warehouses were updated in weekly or monthly batches; most today feature a daily batch update.

But, in the decade ahead, most businesses will want to respond to events within minutes or seconds of their occurrence; on websites and in other venues where the enterprise is interacting with a customer in real time, some businesses already rely on the data warehouse to respond to events in a sub-second time frame.

These scenarios mean that new data must be ingested and visible to queries in a short period of time, typically requiring either that the data warehouse ingest data streams a record at a time in near real time. In some situations, micro-batching may be acceptable. In either case, an ability to accept new data continuously and incorporate it more or less immediately is — or will soon become — a data warehouse requirement in most businesses.

Much of the new data arriving continuously is time-series data. Time-series data is any sequence of records whose primary purpose is to describe how the value of something changes over time. Examples include the price of stock, the vital signs of a medical patient and the temperature of a turbine. There are additional analytical requirements associated with time-series data; there are challenges associated with missing and incorrect sensor readings; and, all the challenges of managing and analyzing time-series data are amplified by the extremely volume and uninterrupted, time-sensitive flow of such data.

D. DATA SCIENCE/ARTIFICIAL INTELLIGENCE/MACHINE LEARNING

Early data warehouses were expected primarily to select relevant data and aggregate it along various dimensions, such as product, customer, store and so on.

The advanced statistical methods of data science and the algorithms associated with machine learning were not initially seen as relating to the data warehouse. In fact, data scientists often rejected the data warehouse as unlikely or unable to manage the data that most interested them: the newer data sources that illuminated customer behavior or product failure. For some uses, data scientists prefer raw data. Raw data was not generally available via the data warehouse, since the practices assumed that it was desirable to cleanse and integrate the data before allowing users to access it.

But now an increasing number of data scientists and machine learning specialists want to leverage the integrated, cleansed business data available in the data warehouse along with the new data sources they often work with in raw form.

Now that it is desirable for the data warehouse to be able to host data science and machine learning processes, it must also enable the much more complex and frequently compute-intensive analytics they require.

In an earlier phase of data warehouse development, the practice was to extract a set of data from the data warehouse; deliver it to another platform dedicated to analytics; and, analyze it there. This pattern of operation must continue to be supported in the decade ahead. However, today's data volumes are so large that the only practical way to apply the analytics in many cases is on the production data in place. Thus, all varieties of advanced analytics and artificial intelligence must be supported directly on the data warehouse platform, a much different requirement than we have seen on the traditional data warehouse.

Purpose and Methodology for this Report

This WinterCorp Research Note describes the Cloudera Data Warehouse and its significance to customers for data management, query and analytics. In developing this report, WinterCorp drew on its own independent research and experience, interviewed Cloudera employees, attended Cloudera events and analyzed Cloudera documentation and literature. Cloudera was provided an opportunity to comment on the paper with respect to facts, in its capacity as the sponsor of this research. WinterCorp has final editorial control over the content of this publication and is solely responsible for any opinions expressed.

WinterCorp

E. SELF-SERVICE

In the modern enterprise, business analysts who work hands on with data increasingly want the capability to bring data they choose — perhaps from an external source — into the data warehouse environment and then join it with production data.

For example, if a business analyst discovered an online data source able to enrich her understanding of customer preferences, she might choose to bring the data into the enterprise, marry it with existing customer data, and launch into an analysis with the goal of bringing new insight to marketing or customer service activities.

The idea that a business analyst could do this was at odds with the traditional philosophies of the data warehouse, because traditionally, (a) only DBAs could introduce new data into the environment; (b) such data was made available to end users only after enterprise data modelling, ETL and other processes; and, (c) the enterprise data warehouse was viewed as carefully fenced production environment that provided no place for end user experimentation.

So, for some time, companies have needed a capability for end users to experiment with combinations of their own data and production business data, in an environment in which they don't create security risks and don't disturb production operations. This type of self-service data lab is a capability needed in the modern data warehouse — and contrasts sharply with the traditional notion in which users have self-service only to the data that DBAs have made available via a strictly controlled process that typically takes months to make a new data source available.

F. SPEED OF BUSINESS

The modern enterprise must compete in a much faster moving world than we had in the early 1990s when data warehousing emerged. Recall that the world wide web only came into commercial use around 1995. eCommerce emerged in the late 1990s, when most of a skeptical public refused to shop online.

For 25 years, the internet, social networking, smart phones, IoT and many other developments have changed our lives and relentlessly increased the speed of business. New companies digitize business in a previously stable market, and whole industries must react in a short time, or be left behind.

The early data warehouse was all about stability, control and the cleanest possible data. The modern enterprise still needs this in certain areas of activity, but in many areas, business can't afford to wait.

So, the modern data warehouse must be agile in a profound sense. It must quickly accommodate change; it must facilitate rapid prototyping and experimentation; and, it must support rapid implementation at scale of new data-based business solutions.

This need for agility and rapid change requires advances in architecture; new product capabilities; and, new and more varied business processes around the data warehouse and its governance.

G. EXTREME SCALE

There have been enormous changes in the data warehouse environment since its emergence in the 1990s, but no single change is more dramatic than the growth in scale.

The largest data warehouse in the world in 1996 contained two terabytes of data. Today, Facebook's data warehouse, conservatively estimated from published reports, contains at least a million times as much data — something more than two exabytes.

There have been many advances in computer hardware that make such a change more practical than it might otherwise seem. For one thing, the capacity of data storage devices have increased very rapidly and a new type of storage — object storage — has emerged. So, it is feasible to store such volumes of data.

But the other aspects of scale: manipulating, selecting, joining and retrieving data on such a massively increased scale is truly a daunting engineering challenge.

The traditional data warehouse has needed to grow at a remarkable pace over these years just to manage the tabular data on which data warehousing was originally founded. But to also accommodate the many times larger and more varied newer types of data has proven to be much larger challenge yet. And that is the challenge of the 2020s: how to manage yet more massive, more varied and more rapidly growing collections of data, somehow providing for them services similar to those provided in the traditional data warehouse.

H. CLOUD

We have covered some of the extraordinary disruptions of the world of information technology since the emergence of the data warehouse 25 years ago. Among those already discussed are the internet, smartphones, big data and IoT. Few people, if any, would have foreseen just how these developments have actually played out and how much they have changed our lives.

But, there is another very big change we have not yet discussed: cloud. The adoption of cloud computing is probably the single biggest change in the information technology field in the last ten years. And, in its modern form, cloud computing was completely unknown when most data warehouse programs were created.

One indication of the impact of cloud computing is that approximately 10% of all data centers were closed in 2017, even as investment in information technology continued to grow at a rapid pace. Cloud revenues are estimated to have topped $70 billion in 2018, an increase of 48% over the prior year.

As more and more systems are developed and operated in the cloud, of course there will be an increasing number of data warehouses implemented in the cloud. So the modern data warehouse needs to provide for the cloud, without doubt. But the requirement is actually more than simply "run in the cloud." Enterprises are buying cloud based business solutions, right and left. Such purchasing is not generally centralized and it would often not be practical to buy such solutions only in a single cloud.

Beyond that, many enterprises are wary of "putting all the eggs" into one basket, fearing vendor lock-in followed by unacceptable business terms or prices. So, some companies are in multiple clouds by choice and others are in multiple clouds by chance.

Regardless of the reason, the data of a modern enterprise will not exist in just one cloud. And, in many companies it will not all be in the cloud. While some companies are, or are planning to be, entirely cloud based others are consciously pursuing either an on-premises or hybrid strategy. While cloud is clearly advantageous in the presence of highly variable or unpredictable workload, it is often not price advantageous when running a steady, round the clock workload.

Think of it this way: if you are hosting a wedding, are you always better off buying the alcohol by the glass? Sometimes you can save a lot of money by buying the alcohol yourself by the bottle or case. In other circumstances, it just makes more sense — and saves money — to pay by the drink.

Cloud computing, in general, is buying by the glass. Cloud vendors provide other options, such as dedicated servers, but sometimes you can still save a lot of money by running your heavier work in your own data center.

And, then there is security. Yes, there has been much progress in cloud security. But, there are still data breaches and some types of businesses are either not heading for the cloud or are doing so very selectively.

What all of these trends add up to is this: the modern data warehouse needs to provide for cloud, multi-cloud, hybrid and on prem architectures. Only if you can advantageously offer all the options can you meet the needs of most enterprises for the decade ahead.

In fact, the modern data warehouse requirement goes beyond simply providing for these deployment options to providing for effective data governance across the multiple environments. So, customers need to be able to find the data relevant to a search or analysis; they need to be able to find the relevant metadata; and, they need to have data governance policies applied to data, regardless of where it is deployed. So, if there is an encryption policy that applies to personally identifiable information (PII) — such as social security number or name and address — then that policy must be implemented regardless of whether the data is on prem on in any of several clouds. The same principle applies to policies regarding data quality, data access, data retention and data backup — all of these are typical data governance policies.

I. TRADITIONAL ENTERPRISE QUERY AND ANALYSIS REQUIREMENTS AT SCALE

We have covered several requirements that have changed drastically for the modern data warehouse, but we are now concluding with something that has outwardly stayed much the same: the need to be able to process structurally complex SQL queries efficiently at scale. This was a difficult engineering challenge in the mid-90s when data warehousing emerged and it is still difficult. It remains difficult because the data is now, as we have discussed above, about a million times larger. Clearly, only a few companies have the data volumes of Facebook, but even brick and mortar enterprises have data volumes many times larger than they did 25 years ago, so they face a similar challenge when their business requires that really complex SQL queries get processed.

What is meant by "structurally complex"? Just this: in a mature enterprise data warehouse that fully represents the major aspects of an enterprise in one integrated data store there are typically multiple large tables. And, unlike the simplified "star schema" examples used in textbooks , these multiple large tables often have a complex web of relationships among them. That means that they do not always get joined along the same path or on the same key.

So, for example, in health care, the most common join is on patient ID. You might imagine you can organize the data warehouse around that. But that is not the only join of interest. Users will also want to join the data on disease, treatment, provider, test results, medications and other topics. Sometimes the join will be on patient ID, but sometimes not.

So, for real freedom in analyzing just your structured data via SQL, there has been a need from the beginning to be able to perform complex and unpredictable queries involving a variety of joins on large tables. And, here's a secret: some popular data warehouse products could never do this well. Others could. And this has limited the growth and success of many data warehouse programs.

Looking ahead into the coming decade, I believe this capability will be more important than ever. The complex queries involving the varied and structurally challenging joins will become more important, not less. And, now, many will involve semi-structured data as well as the traditional tabular data. And, now, the tables are much, much larger. So, this is a traditionally demanding requirement which is many times intensified in the modern data warehouse.

3 A Modern Approach to Data Warehousing

So, you have just read my description of the requirements of data warehousing in the modern enterprise in the 2020s. How is it possible to meet those requirements?

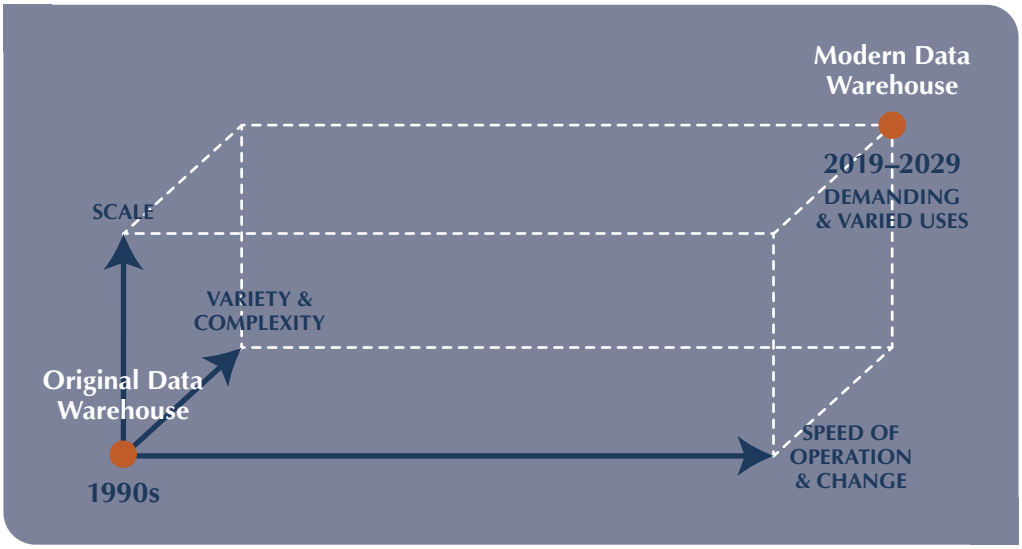

Figure 1: How Data Warehousing Must Change

Basically, there are three parts to the puzzle. You need:

- a modern conception of what the data warehouse is and how it is organized;
- a modern data warehouse platform; and,
- a set of modern data warehouse practices.

You need all three to meet the requirements described in *Section 2*, above.

3.1 THE CONCEPT OF THE MODERN DATA WAREHOUSE

In the modern data warehouse, new capabilities and extended capabilities are provided in the following categories:

Data Management

- Any volume of data can be stored
- Any type of digital data can be stored
- All stored data can be queried and retrieved
- All data can be protected against loss due to hardware failure, software failure or human error

Data Curation

- Full metadata and lineage is captured and maintained for all data; all data can be found via a universal data catalog
- Data is curated at a level appropriate to its intended use; in general, a copy of the data originally submitted in raw form is retained; data is retained in three or more zones, typically including a raw zone, a curated zone, and a published zone

- While the core enterprise data may be organized into a single, integrated, enterprise data model as is commonly implemented in the classic EDW, the modern data warehouse may also have multiple separate collections of data (sometimes called data marts), each of which may have its own model, either explicit or implicit in its data definitions
- Schema on read is optionally supported; thus, some data may be ingested into the modern data warehouse as undefined, raw data, to which schemas are to be applied upon use
- New data sources are readily added
- Schema changes in existing data sources are readily implemented
- Schema changes on stored data, including curated data, are readily implemented, in a largely or wholly automated fashion

Deployment

- Data can be deployed on prem, in the cloud, hybrid cloud (cloud + on prem) and/or in multiple clouds
- Data can be stored on a storage medium selected on the basis of cost effectiveness, access time and other relevant considerations, determined where appropriate by policies automatically applied

Security

- Any data can be secured with mechanisms appropriate to its sensitivity, determined where appropriate by policies automatically applied

Queries, Analytics, Models, AI/ML

- Business intelligence and data science tools can readily be used to access the data, with appropriate performance, regardless of data volume
- Any data can be analyzed with any variety of analytic
- Time series data is readily queried and analyzed
- Data is readily queried and analyzed with respect to geospatial attributes
- Data is readily available for model development and training of machine learning applications
- Data is readily available for production use of models and machine learning applications
- Traditional data warehouse queries and analytics are readily supported at scale

Service Levels/Workloads

- Multiple workloads with different service level requirements can be run concurrently, with each managed according to its requirements

End User Access

- End user self service: end users can readily perform straightforward tasks without coding, including bringing new data into the modern data warehouse, within a segregated data laboratory environment, and cleansing it, transforming it in simple ways, and combining it with existing enterprise data

Data Virtualization

- Data maintained in external databases and data lakes can be incorporated into the data warehouse virtually, for purposes of query and analysis

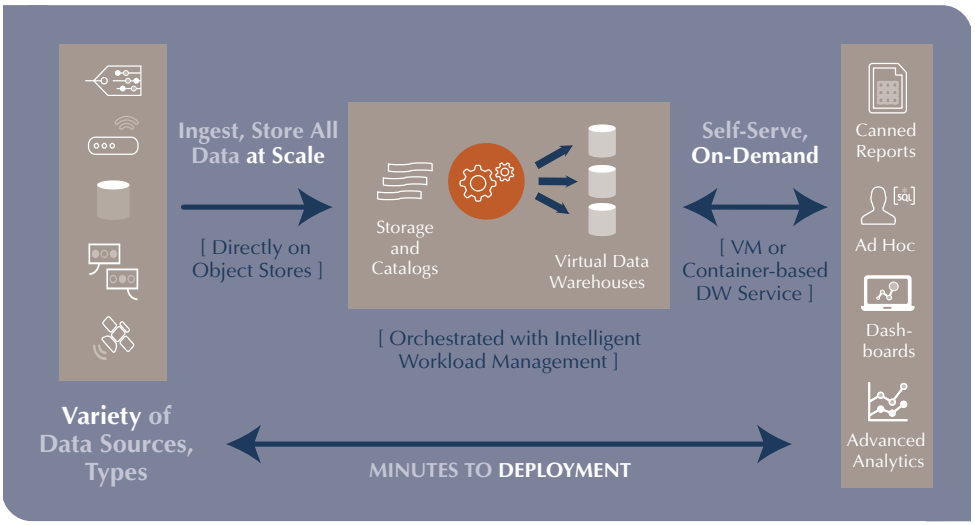

Figure 2: The Modern Data Warehouse (Source: Cloudera)

Modern Data Warehouse Platform. In the vision of the modern data warehouse presented just above, all of the limitations of the traditional data warehouse are overcome to some significant degree. This is an ideal, and any real data platform will have some limitations.

However, to qualify as a modern data warehouse, a product will need to be significantly better than earlier systems concerning most or all of the described capabilities.

Therefore, the modern data warehouse platform will need to feature the performance and scalability to deliver at required service levels for very large volumes of data. It will need the flexibility to handle great data variety and great analytic variety. It will need the architecture to be able to deliver these services in the cloud, on prem or in hybrid configurations. It will need the agility to enable very rapid time to value while responding to a wide variety of new requirements.

As with any other product category, it will be in the tradeoffs and engineering choices that a given modern data warehouse will emerge as the best choice for a particular set of uses.

Modern data warehouse processes. Note that some of the capabilities listed above require changed or new work processes in addition to system capabilities.

For example, if an end user can introduce new data into the data lab in order to experiment with it and assess its value, how then does this data become incorporated into the data warehouse and made into a data asset that can be shared across the enterprise?

How does this get accomplished without the extensive delays emblematic of traditional concept? The answer lies in unbundling the concepts of data sharing; data preparation; data modelling; and, enterprise data management.

So, for the new data source to end up as part of the core enterprise data model — integrated with all other data to which it is related and curated to the highest standards for quality — well, that will still be an extensive, multi-step process.

But, not every data source requires this level of curation and integration. And, certainly not for initial uses by small work groups.

So, the modern data warehouse will have multiple levels of curation, for different cases of sharing, use, monetization, compliance, etc.

For the uses requiring less rigor, the processes can be simpler and more streamlined. And, for private use by a single user in the data lab, the processes should be minimal (perhaps only that the external

data undergo a virus scan, an automated compliance scan, and get logged). In general, the more widespread the use of the data and the greater its significance to the major and/or more critical processes of the business, the higher the standard of curation that must be applied.

Data governance and processes will be different for each level of curation. There will also be processes for promotion from one level to the next, and it will be critical to tag data with its level of curation and maintain its history with respect to all curation processes.

With this approach it will be possible to enable rapid incorporation of new data at lower levels of curation and rapid development of new business solutions that do not depend, at least initially, on highly curated data.

4 The Cloudera Data Warehouse

Cloudera has recently introduced the Cloudera Data Warehouse, positioning it as a modern data warehouse in the sense described in this report.

The Cloudera Data Warehouse provides capabilities for storing, managing, retrieving and analyzing data of extremely large scale and great variety. It can be deployed on a Hadoop cluster or, in the cloud — in fact, it can be deployed in any of the popular cloud services. The Cloudera Data Warehouse can also be deployed in a hybrid (on prem plus cloud) configuration. The Cloudera Data Warehouse supports multiple levels of curation and provides facilities for data science and machine learning, in addition to supporting tools for business intelligence and traditional data warehouse analytics. The Cloudera Data Warehouse provides support for time series data, geospatial data and a wide variety of other data types and formats. It provides facilities to secure data against unauthorized access and protect data against loss due to hardware failure.

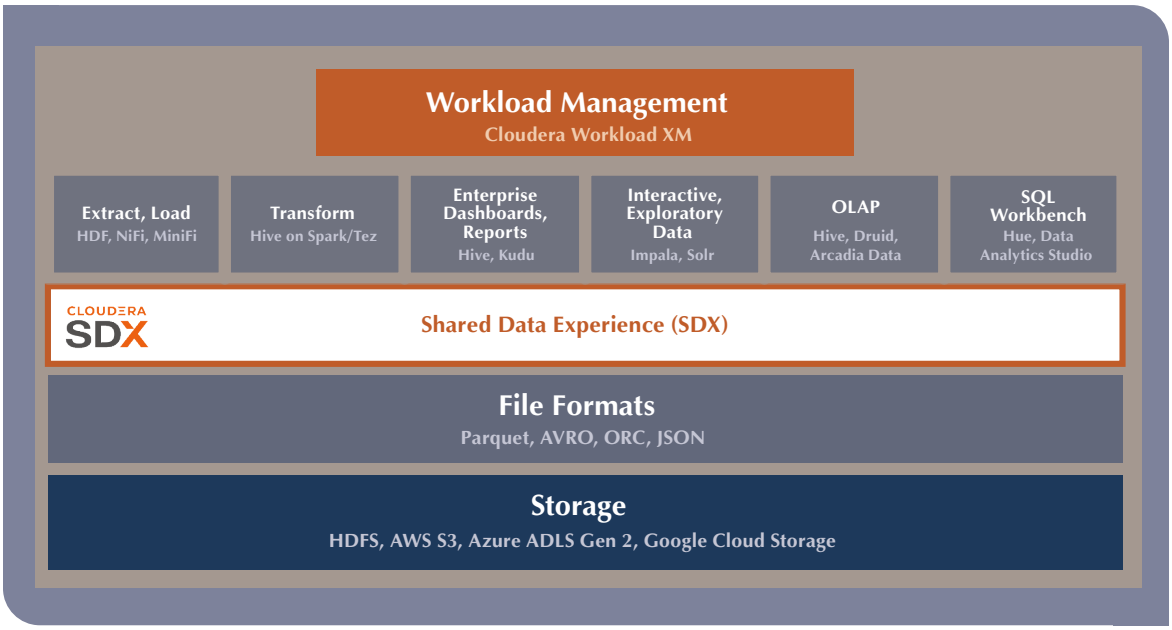

Figure 3: Cloudera's Modern Data Warehouse (Source: Cloudera)

The Cloudera Data Warehouse provides most of the facilities of the idealized modern data warehouse as described in *Section 2*. In addition, the Cloudera Data Warehouse is focused on aspects of the concept likely to be of growing importance in the years ahead: large scale; higher data variety; higher analytic variety; and, multiple levels of data curation.

4.1 CLOUDERA'S SHARED DATA EXPERIENCE (SDX)

The foundation underlying the Cloudera Data Warehouse and the wider Cloudera Data Platform is Cloudera SDX, a powerful data fabric for complete data security, governance and control across infrastructures, providing ultimate deployment choice and flexibility.

Cloudera SDX enables data and metadata security and governance policies to be set once and automatically enforced across data analytics in hybrid and multi-clouds. As a result, self-service discovery and access of both data and analytics can be provided in a safe and secure manner, across infrastructures, increasing business insight and agility as well as reducing IT overheads and security risk without being locked into any one infrastructure provider.

4.2 DATA WAREHOUSE FUNDAMENTALS

What does the Cloudera product line have to do with data warehousing? Five years ago, this author would have said, "Not much." But Hive, Impala and some other key components of the Cloudera Data Warehouse have matured greatly in that period.

In particular, Hive managed tables have grown into something that provides key strengths of the relational database — and this has been accomplished without compromising important aspects of the flexibility that Hive has always provided.

As a result of recent developments in Hive, and culminating in Hive 3, introduced in late 2018, Cloudera now provides:

- Updates and deletes on Hive tables, which previously supported only the appending of new data;
- Optional ACID properties – the ability to have the system to enforce transaction semantics, an important guarantee for the integrity of data undergoing any type of change in a database[1];
- Support for the definition of a primary key;
- Support for constraints, including primary key/foreign key constraints (to be released shortly);
- A much enhanced, cost based query optimizer, able to select the best available plan for the query and, re-optimize when query execution doesn't go as expected;
- Materialized views: virtual tables in which data can be pre-selected and pre-joined, accelerating query performance; and,
- Data analytics studio: a desktop tool for understanding query performance and managing query execution, among other useful capabilities.

These are in addition to the large and fundamental advances in Hive 2.0, known informally as Hive LLAP (Live Long and Process). LLAP provides a hybrid execution model that consists of a separate long-lived process, called a daemon, which replaces direct interactions with the HDFS DataNode, and a tightly integrated DAG-based framework. Caching, pre-fetching, some query processing and access control are moved into the daemon. Small/short queries are largely processed by this daemon directly, while any heavy lifting will be performed in standard YARN containers. The effect is a large gain in efficiency for both short and long queries

There are many other improvements in Hive over the last few years — far too many to cover in this report. But, this list alone illustrates how much Hive has grown much more capable of supporting the data warehouse.

In addition, the ORC columnar data format, long supported in Hive, is going to be supported as a native data format in Impala, making it easier for customers to use Hive and Impala together in a data warehouse solution architecture.

[1]ACID properties are supported for single-table transactions.

Druid. Apache Druid provides fast analytical queries, at high concurrency, on event-driven data. Druid can instantaneously ingest streaming data and provide sub-second queries to power interactive UIs.

In the Cloudera Data Warehouse, Druid can be used together with Hive to address a wider range of query needs. Druid builds OLAP cubes — that is, summaries of data along common dimensions such as time, geography, product, price, etc.

For example, if a company is running a world wide advertising campaign and analyzing buying behavior and social media impact by region, an analyst might want to compare sales this month in India with sales last month. Similarly, it may be of interest to compare sales in India by week with sales in South Africa. Such queries can be processed very rapidly by Druid, using its pre-built aggregates.

As a result of recent developments, both Hive tables and Druid cubes are available via a unified SQL interface in the Cloudera Data Warehouse.

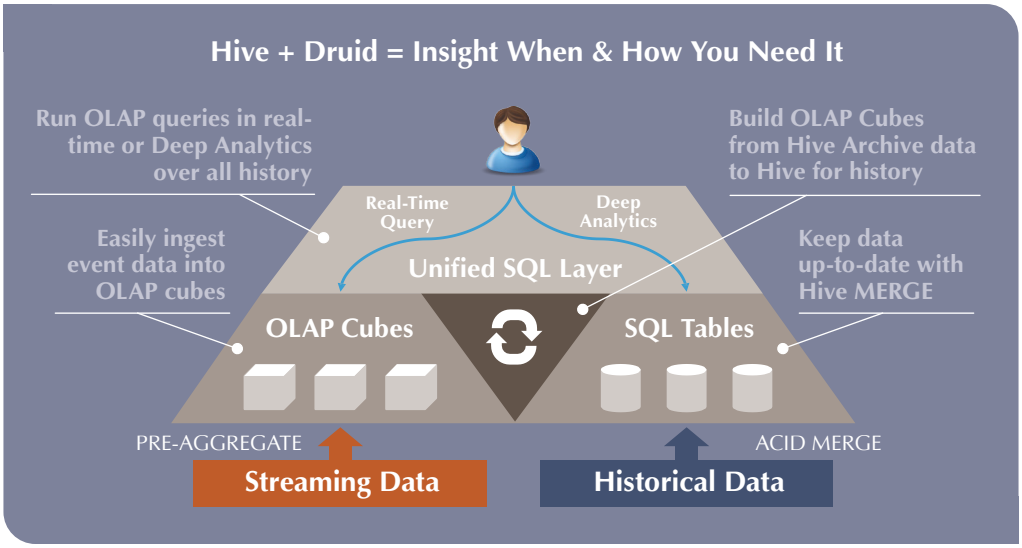

Figure 4: Hive & Druid Accessed via a Unified SQL Layer

The connection of Hive and Druid is one example of a Hive Warehouse Connector. Via Hive Warehouse Connectors, Hive queries can also access data in Kafka queues — that is data which has streamed into the data warehouse and has not yet been stored in a Hive table.

In another example, Spark developers employ a Hive Warehouse Connector so that Spark applications can access data in managed Hive tables. In this case, the use of the connector enables Hive to retain control of the data so that it can enforce constraints and transaction semantics. This is a key aspect of the recent advances that make Hive to manage data with the integrity expected in a data warehouse.

4.3 WORKLOAD MANAGEMENT

One of the fundamental goals of a data warehouse is to enable sharing of data across the enterprise. The underlying idea is that business data is a valuable asset and its value is fully leveraged only when those authorized to use it are able to maintain it and access it whenever that may be necessary. But, in an enterprise employing or servicing many people, it is inevitable that multiple processing will be operating at the same time. Some of those processes will be loading or updating data; some will be long-running jobs doing large scale analysis or reporting; others will be supporting interactive users with small queries that must be satisfied quickly.

Recently Cloudera has introduced its Workload Experience Manager (Workload XM). Workload XM is a tool that provides insights to help you gain in-depth understanding of the workloads you send

to clusters managed by Cloudera Manager. In addition, it provides information that can be used for troubleshooting failed jobs and optimizing slow jobs that run on those clusters. After a job ends, information about job execution is sent to Workload XM with the Telemetry Publisher, a role in the Cloudera Manager Management Service.

Workload XM uses the information to display metrics about the performance of a job. Additionally, Workload XM compares the current run of a job to previous runs of the same job by creating baselines. You can use the knowledge gained from this information to identify and address abnormal or degraded performance or potential performance improvements.

4.4 TIME SERIES DATA

The storage and analysis of time series data has often been difficult to accommodate in the traditional data warehouse product. A time series is any sequence of records whose primary purpose is to describe how the value of something changes over time.

Time series data is present today in many industries from Wall Street (the list of all changes in the price of a security is a time series) to medicine (the vital signs of a patient and how they change over time) to energy production (the moment by moment changes in the amount of energy produced by a turbine). Time series data has been around for a long time, but it has been awkward to deal with in the data warehouses in use in most companies.

There are three main problems:

- Extreme data volumes;
- Extreme data variety;
- Unusual and complex analytics.

As an example of volume, consider that one energy company is concerned with the data flowing out of more than 50 million smart meters.

As an example of variety, consider that a large industrial vehicle may have thousands of sensors, where the data flowing out of each sensor may be formatted differently. In a traditional relational database, tables are expected to have a fixed format. If you put the data from each sensor in a different table, you have so many tables, that it becomes extremely clumsy to do analysis.

The analysis to be done is complicated by missing values: sometimes sensor readings are lost because the device has a transitory malfunction; sometimes there is a network error. The analysis is about recognizing patterns and trends, sometimes tricky to discern in noisy data. In terms of a relational table, these are patterns that occur down the columns, not along the rows. The analysis involves predictive analytics: when will this part fail? What is the latest good time to replace it?

Cloudera Data Warehouse has the capabilities needed to store and analyze time series data and Cloudera customers are doing it on a very large scale. One telecom has implemented a system in which five million events per second are tracked. The database stores 120 TB of data per day. 138,000 different metrics are tracked. Queries must be processed within five seconds. In implementing this system on Cloudera Data Warehouse, the customer has reduced system operating expense by 86%.

While time series data has been around for a long time, it has often been impractical to keep it in a data warehouse, because the most widely used products either could not perform at large scale or could not store the data economically. This has often meant that the data would either reside in an application-specific silo or would be lost. Now there is an opportunity to manage the time series data in a data warehouse and treat it as an asset, which opens up the possibility of extracting much more value from it.

Time series data is one example of many types of specialized data which have not been stored in data warehouses in the past. The “big data” era which emerged about ten years ago has brought to light many types of specialized scientific, engineering, imaging, audio and other data. With the

modern data warehouse we now have an opportunity to treat all these varieties of data as assets; to make them accessible; to make it easier to query and analyze them; and, to make it easier to integrate them with other types of data in the enterprise.

4.5 CLOUD

Cloudera Data Warehouse runs on premises on Hadoop and is also available on major cloud services. Further applications and databases are portable among instances of the Cloudera Data Warehouse, whether they are running on different clouds or on premises.

5 Customer Examples

Customers have been building data warehouses with the Cloudera product line for several years. They did this even before Cloudera formally introduced the Cloudera Data Warehouse offering.

A bank engaged in fraud prevention has a Cloudera Data Warehouse with 500 TB of data in 167 databases against which they run two million queries a day. Over 240 users access the data.

A global pharma has a Cloudera Data Warehouse for new product development with 8 petabytes of data (PB) against which they have implemented over 200 use cases. 8700 users share this data, which was formerly stored in 200 separate systems.

Figure 6: Telco Data Capture and Analytics (Source: cloudera.com)

A telco has a Cloudera Data Warehouse for business optimization. Their system has over 100 million data sets used by some two million users. One extraordinary aspect of this system is the response time: queries must be satisfied within one second.

In another example, Globe Telecom in the Philippines was experiencing very rapid growth in their data network, to the point where it grew 66% in one year, generating a total of over 600 PB of data. Revenues were not growing as fast as data, so they needed to get insight into what was happening and rapidly develop a plan. Using Cloudera Data Warehouse, they implemented Customer 360, so as to have a more complete understanding of what customers were doing and how this related to their spending. With the Cloudera solution they were able to manage the enormous data volumes

affordably; they were able to complete analysis much more rapidly; and, they were now able to segment their customer base in a way that was relevant to their business situation. This enabled them to better align mobile data consumption with service offerings and pricing.

This example combines several features of the modern data warehouse as it is defined in this report: very large scale, very rapid growth, a pressing need for insight, a need for affordable solutions. Globe Telecom experienced a disruption in their business and needed a set of capabilities that their previous data infrastructure could not provide. They took a modern data warehouse approach and were able to move forward with a business solution.

Hundreds of substantial Cloudera data warehouses are in production today, according to the company. These range in size from a few terabytes (TB) to thousands of terabytes, also known as Petabytes (PB). These demonstrate that a variety of modern data warehouses have already been created with Cloudera products and that customers are realizing value from them

6 Conclusions & Recommendations

On the basis of its independent research, WinterCorp recommends that companies investing in a modern data warehouse at enterprise scale look closely at the Cloudera Data Warehouse.

Among the large scale data warehouse solutions on the market today, Cloudera is unique in offering an open source approach, enhanced by a line of commercial grade software products designed for enterprise scale. The Cloudera Data Warehouse is available on premises and in leading cloud services. It is thus one of the few data warehouse architectures proven in large scale use with which the customer can build solutions that are portable across clouds and across the enterprise fire wall.

Cloudera says that it has 1200 customers running the Cloudera Data Warehouse in production, resulting in annual revenues of about $200 million.

Examples of impressive customer implementations are described in this report.

A word of caution, however, is in order. The modern data warehouse, as it is defined in this report, represents a truly enormous variety of requirements for data management, query performance and analytics. No one product will be best for every requirement. Therefore, though WinterCorp recommends that enterprise customers take a look at the Cloudera Data Warehouse, we also recommend careful evaluation against a company-specific set of strategic data warehouse requirements. Only with a such a careful, strategic evaluation can a foundation for success be created. ●

WinterCorp is an independent consulting firm expert in the architecture and strategy of the modern analytic data ecosystem.

Since our founding in 1992, we have architected and engineered solutions to some of the toughest and most demanding analytic data challenges, worldwide.

We help customers define their data-related business interests; develop their data strategies and architectures; select their data platforms; and, engineer their solutions to optimize business value.

Our customers, with our help, create and implement cloud, multi-cloud and hybrid cloud architectures; they create the data foundation needed for data science, artificial intelligence and machine learning.

Our customers get business results with analytics in which their return is often ten or more times their investment.

When needed, we create and conduct benchmarks, proofs-of-concept, pilot programs and system engineering studies that help our clients manage profound technical risks, control costs and reach business goals.

We're expert with structured data, unstructured data, and semi-structured data — with the products, tools and technologies of data management for data analytics in all its major forms.

With our in-depth knowledge and experience, we deliver unmatched insight into the issues that impede scalability and into the technologies and practices that enable business success.

WinterCorp
www.wintercorp.com
42 DERBY LANE, TYNGSBORO, MA
617-695-1800