cloudera
Ask Bigger Questions

WHITE PAPER

USING CLOUDERA TO IMPROVE DATA PROCESSING

Table of Contents

What is Data Processing?

Data processing is critical to supporting organizations' everyday operations such as generating reports for suppliers and customers, measuring internal metrics day to day, and reporting quarterly financial results. Effective data processing operations enable companies to efficiently create products that revolutionize how people communicate, shop, manage their finances, and learn about the world. Leading organizations have mastered the art of collecting data and bringing it into a system that can store and make data accessible to the enterprise for increased operational efficiency.

For example, telecommunications firms collect detailed records about how each call is initiated and routed in order to provide accurate bills and to ensure high service quality. Online retailers track what consumers browse and buy to make smart inventory decisions and to guarantee high quality and timely shipments. Financial firms combine market data with individuals' transaction details so they can be certain that money is being managed and transferred correctly.

Raw data is generated from sources like call detail records (CDR), point of sale (POS) transactions, web browsing and clicks, electronic medical records (EMR), financial trades, and data management tools such as enterprise resource planning (ERP) systems. In traditional IT infrastructures, an extract, transform, load (ETL) process is employed to collect raw data from various sources, transform it into a consistent format that can be understood by relational database management systems (RDBMS), and then load it into the database or data warehouse for storage and analysis.

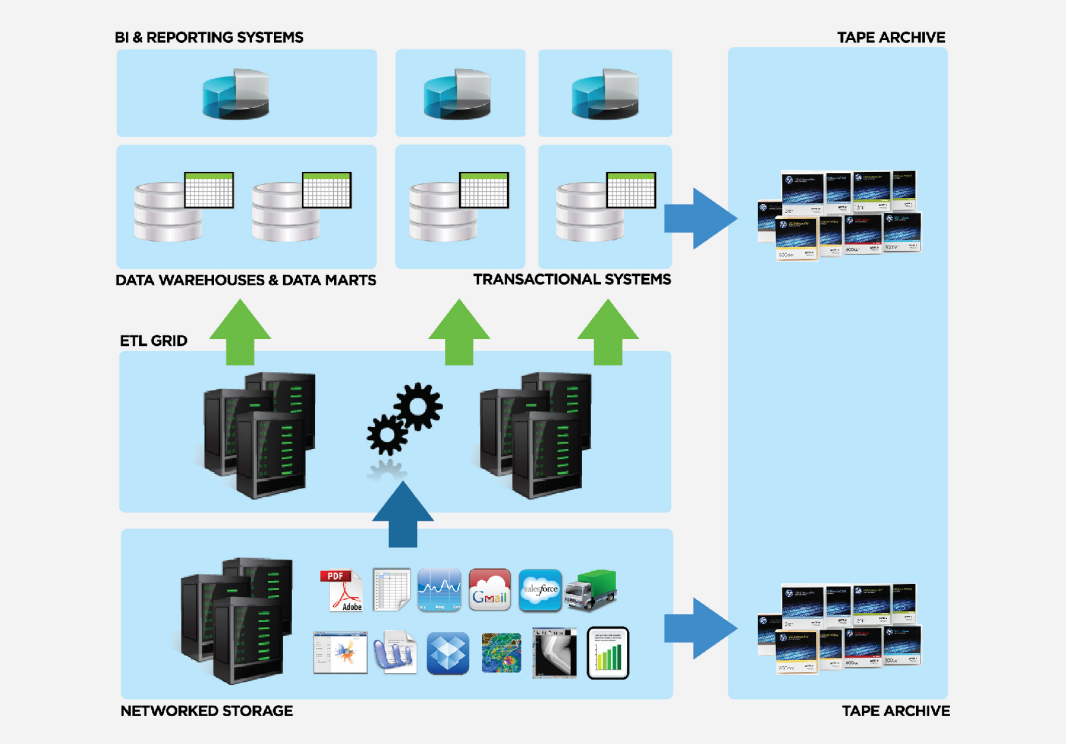

Diagram: Traditional data management structure.

A recent survey of large financial services firms, telecommunications carriers and retailers indicated that storing data in an RDBMS typically runs between $30,000 and $100,000 (USD) per TB per year in total costs.

CHALLENGES

Flexibility and Data Quality Constraints

Raw data is not perfect: source systems can be misconfigured, reporting formats change and third party data sources may contain mistakes. Traditional data quality processes look at individual records using fixed rules for identifying errors, and the errors are kicked out or corrected according to rules after which processing continues. These data quality checks are static and fragile as data changes. When a valid change is introduced, it requires the entire data pipeline to be updated and tested before it can be rolled out. A small mismatch in the source system, transformation pipeline, or schema mapping would potentially cause every record to be rejected. Having a rigid schema that is required from data source through to target means that everything must always line up perfectly. This is a critical constraint that transactional systems have, and it breaks down when looking at new data sources that change rapidly and vary widely.

Too Much Time

While limiting flexibility and hindering data quality, the systems in place to extract, transform and load that data into the warehouse have become a bottleneck as data volumes have exploded. Because unstructured data must be reformatted to fit into a relational schema before it can be loaded into the system, it requires an extra data processing step that slows ingestion, creates latency and eliminates elements of the data that could become important down the road.

Meanwhile, the business is demanding fresher data encompassing even greater history to fuel daily decisions so they can keep up with today's fast-paced economy. In one case, a national telecommunications carrier found themselves struggling to keep up with billing related processing times which were starting to take days to complete. Their raw and archived data were filling up the enterprise data warehouse (EDW), which was not designed to run a heavy processing workload. Billing statement processing is critical to their business, and the raw data they collect has the potential to unlock transformational value about detailed end user interactions.

Too Much Data

The data processing infrastructure that was developed to run business operations over the past decades is having trouble keeping pace with today's digital landscape. Across every industry, daily interactions and transactions are moving online and business services are becoming increasingly automated. Volumes of multi-structured, machine-generated data from a variety of sources have skyrocketed, and smart companies want to capture and make use of it all. As data volumes grow and the complexity of data types and sources increases, data processing workloads take longer to run and the time available for reporting and analysis is reduced.

Too Costly

A recent survey of large financial services firms, telecommunications carriers and retailers indicated that storing data in an RDBMS typically runs between $30,000 and $100,000 (USD) per terabyte (TB) per year in total costs. Due to the considerable growth in data generated and managed, most large enterprises employ a "multi-temperature data management" strategy to reduce costs. In this type of infrastructure, "hot data" that is frequently accessed for reporting and analysis is kept in the data warehouse. "Warm" or "cold data" — which is typically older, considered less valuable, and accessed less frequently — is offloaded to a data storage system that offers cheaper storage and is not readily available for reporting or analysis.

Summary of Challenges

The timeliness and cost of processing data has begun to exceed the capabilities of existing data management systems. Homegrown implementations run into challenges when trying to scale to meet the data and processing needs of the business. Databases are a costly utility for data processing, and organizations are constantly trying to maximize their investments in relational analytics technology.

Organizations are constantly trying to maximize their investments in relational analytics technology.

SOLUTION

Flexible, Scalable, Reliable Data Processing at Low Cost

Meet Hadoop. Developed by Cloudera architect Doug Cutting, Hadoop is open source software that enables distributed parallel processing of huge amounts of multi-structured data across industry standard servers. Hadoop has a number of capabilities that make it an excellent platform for data processing:

- **Flexibility:** The underlying storage in Hadoop is a flexible file system that can hold any type of data. Because of this, Hadoop can accept any kind of file in any format and store this raw data in perpetuity. Hadoop supports pluggable serialization so it can be used to efficiently and reliably store raw data in its original format. As a result, if there are changes to formats or if data needs to be reprocessed, the original data is still available and does not need to be retrieved from the source systems. This process of formatting the data at query time is known as schema on read and is one of the primary advantages Hadoop offers over traditional systems that require data to be formatted to a schema on write.

- **Scalability:** Hadoop leverages a scale-out architecture that marries self-healing, high-bandwidth clustered storage (via the Hadoop Distributed File System, or HDFS) with fault-tolerant distributed processing (MapReduce). HDFS is structured much like a regular network file system and is optimized for large scale data collection and processing. It is linearly scalable — processing power and storage capacity scale linearly as additional industry standard servers are incorporated. HDFS and MapReduce are both designed to run on scale out systems. The data loaded into HDFS is partitioned sequentially across any number of standard servers with local disk. This makes provisioning new servers very efficient and requires no additional effort on the part of the developer to handle massive changes in data volume. The same MapReduce code that runs on 10GB of data can run on 10PB of data. All Hadoop requires is additional servers that contain both data storage and compute. No data is too big.

- **Data Integrity:** The Hadoop processing framework, MapReduce, offers a robust application program interface (API) that allows both record-at-a-time processing for basic quality checks and single-pass, multi-group processing for efficient execution of more robust quality checks across records. MapReduce includes built-in mechanisms for handling data errors and for detecting large numbers of errors (such as when a data set has changed). This early detection gives operators a chance to identify and quickly resolve data processing challenges.

- **Affordability:** Because Hadoop uses industry standard hardware, the cost per terabyte of storage is, on average, 10x cheaper than a traditional relational data warehouse system. Hadoop uses standard servers with local storage, thereby optimizing for high I/O workloads. Servers are connected using standard gigabit and 10 gigabit networking, which lowers overall system cost and still allows near limitless storage and processing by scaling out. Through its use of local storage and standard networking, Hadoop reduces the total hardware requirements since a single cluster provides both storage and processing.

While Hadoop offers cost-effective data storage and high performance parallel processing of multi-structured data at petabyte scale, the rapidly evolving platform and tools designed to support and enable it are very complex and can be difficult to deploy in production. Leading organizations are working with to Cloudera to improve the operational efficiency of their data processing environments. The Cloudera Enterprise big data platform consists of three things:

- **CDH**, the most widely adopted distribution that bundles popular open source projects in the Apache Hadoop stack into a single, integrated package with steady and reliable releases. CDH is 100% open source.
- **Cloudera Manager**, an end-to-end management application for CDH that simplifies Hadoop deployment and empowers users to improve cluster performance, enhance quality of service, increase compliance and reduce administrative costs.
- **Expert Support**, from Cloudera's highly trained engineering team, which includes contributors and committers to the various open-source projects that are part of the Hadoop ecosystem.

One telecommunications carrier reduced processing times four-fold and reduced processing costs by 91% per year by moving data to CDH.

RETURN ON INVESTMENT

In order to increase utilization of their EDW for reporting and analytics, the aforementioned carrier decided to offload their raw data processing to Cloudera Enterprise. Starting with one petabyte (PB) of data and moving all new incoming data to CDH, this telecommunications company was able to free up precious resources on their EDW, reduce processing times four-fold, and lower data processing costs by 91% per year.

The carrier's previous data processing environment was costing $59 million (USD) each year to manage 1PB of data, broken down as follows:

- $2 million (USD) per year = storage for 1PB raw archive data on network-attached storage (NAS) at $2,000 per TB per year
- $55 million (USD) per year = management and backup of 1PB processed data on EDW at $55,000 per TB per year
- $2 million (USD) per year = administration costs calculated at $1,000 per TB per year

Calculating costs for moving data processing onto Cloudera, the carrier reduced infrastructure costs to $5.1 million (USD) total:

- $5 million (USD) per year = hardware, software and infrastructure for 1PB at $5,000 per TB per year
- $100,000 (USD) per year = administration costs calculated at $100 per TB per year

Additional benefits: increased capacity and availability on the EDW. Previously struggling to manage demands for access to the EDW, by deploying Cloudera Enterprise the carrier was able to offload data processing and historical data storage, freeing up precious space and compute capacity for their reports and analytics.

There are three main tools that streamline data movement from source systems to Hadoop: Sqoop, Flume, and HttpFS.

A CLOSER LOOK: TECHNICAL DETAILS

CDH Data Processing Pipeline and Components

Let's examine a Hadoop data processing pipeline and its components in greater detail. There are three main tools — all components of the CDH stack — that streamline data movement from source systems to Hadoop:

- **Apache Sqoop** is a component of CDH that connects to relational databases. Sqoop allows bi-directional data movement between any component in CDH and virtually any relational database system as well as some NoSQL systems. There are high performance connectors available for a number of EDW systems.
- **Apache Flume** enables any streaming source such as a web or application server, network device, or operational system to load real-time data directly into HDFS, Hive or HBase. Flume supports a variety of source protocols, such as syslog, and can handle any delimited file format. Flume also includes the ability to transform and direct messages on the wire. The result: data from a wide variety of sources can be collected and stored in any format within CDH.
- **Apache Hadoop HttpFS** is a service that provides Hypertext Transfer Protocol (HTTP) access to HDFS for Representational State Transfer (REST) based file movement. HttpFS makes it trivial to integrate CDH into a service oriented architecture. And since HttpFS supports full security integration with the entire CDH stack as well as Secure Sockets Layer (SSL) for on-the-wire encryption, HttpFS provides a fully secured gateway for sensitive data such as customer financial records.

Whether collected by Sqoop, Flume, or HttpFS, all data is fed into HDFS where it lands in an incoming directory structure based on the source system, the type of data stored in the file, and the date and time. Once in HDFS, data transformations may be facilitated by Apache Oozie, which provides rich workflow automation capabilities. Oozie can also handle failures and incremental restarts.

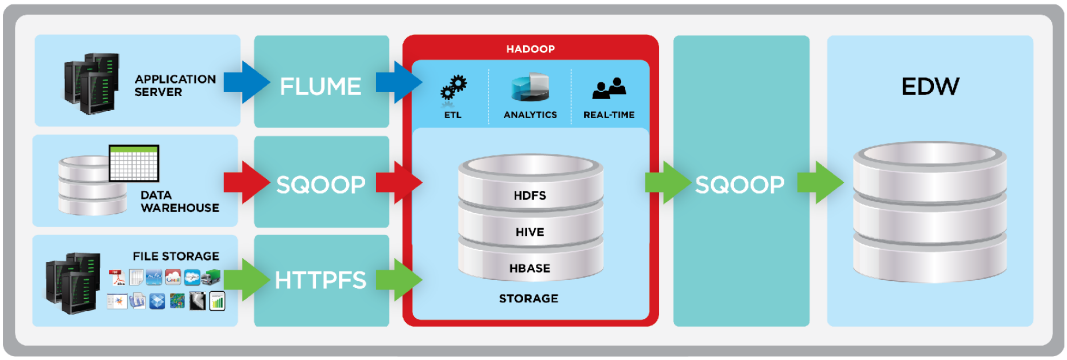

Diagram: Data movement within a typical Hadoop environment.

MapReduce is not always the most efficient abstraction for processing data in Hadoop.

Transformation Development

The underlying facilities in Hadoop provide scalable data storage and primitives for data processing. While any type of data transformation can be written in MapReduce, this is not always the most efficient abstraction for processing data. Today, most transformations are authored in Apache Hive, Apache Pig, or third party commercial solutions.

- **Apache Hive** is an SQL like language, officially known as Hive Query Language or HQL. HQL is accessible to the SQL developer, making it organizationally efficient for people with existing SQL skills to write data transformations in Hadoop. Large organizations who have existing transformations written in SQL stored procedures have found that many translate directly into HQL or can be implemented using HQL and Hive user defined functions (UDFs).

  The primary advantage of Hive over strictly relational database systems: Hive compiles into MapReduce and runs on HDFS. This means that it inherits all of the flexibility, data integrity and scalability properties of Hadoop. Hive can elegantly deal with data format errors, with data that has very intricate and complex defined structures, and with data that is not stored in well-defined data types.

  Hive also allows users to create multiple tables with different logical schemata on the same data sets. Because the underlying storage is all raw data, different schemata can be mapped on to the same data sets without moving or reformatting data. Since the data is being read and processed at query time during data transformations, there is little overhead to also mapping the schema at query time.

- **Apache Pig** is a data flow system with a language (Pig Latin) designed for expressing data processing logic in a sequential fashion. While foreign to SQL developers, it comes naturally to those who write Python or Perl based data transformation code, or to those who write in stored procedures such as PL-SQL.

  Pig also compiles into underlying MapReduce operations; it is flexible and scalable because it runs natively on Hadoop. For transformations that logically flow in sequential steps rather than set transformations, Pig Latin is the preferred data processing developer language for Hadoop.

- **Third party commercial solutions** are also available for building data transformations on Hadoop. Some of the leading vendors in this space are: Datameer, Informatica, Pentaho, and Talend.

Most data transformations are extremely input/output (I/O) intensive.

Infrastructure Integration

Hadoop is a flexible system that can be deployed on a wide variety of servers. For data processing workloads, the typical underlying infrastructure is a storage heavy cluster. These systems typically include twelve disks of 2-3TB per disk, offering 36TB of raw storage per server. The servers are configured with a disk to CPU ratio of 1:1 using dual hex core processors and with a CPU to RAM ratio of 4:1 using 48GB of RAM. The large storage capacity allows organizations to not only capture and transform new data sets but also to store massive volumes of historical data.

Most data transformations are extremely input/output (I/O) intensive. All new data is read, repeatedly combined and processed to produce a resulting data set. Cleansed data sets can be copied to the relational database or warehouse using Sqoop, allowing business and IT users to access their big data using standard reporting and analysis tools. Data from Hadoop is also often brought into the RDBMS for drill down and trending analytics. These database systems already exist in most organizations; common databases or data warehouses in use today are: IBM DB2, IBM Netezza, Microsoft SQL Server, Oracle, and Teradata.

Direct access to Hadoop from BI tools creates opportunities for the broader business community to work with all of their detailed data.

Reporting System Integration

When data processing workflows are complete, the output data is used to run day-to-day business operations. This includes reporting for internal purposes and providing data sets and reports to suppliers and customers. Sqoop is commonly used at the output of a data flow to push processed data to existing relational reporting systems. Since the data output has been quality checked and processed, it can be efficiently loaded, stored and queried from relational databases. Sqoop supports optimized connectors for a variety of RDBMS including Oracle, Teradata and IBM Netezza.

Increasingly, organizations that deploy Hadoop for data processing are also accessing the system directly for reporting. Cloudera includes ODBC drivers for connecting to Hive from traditional business intelligence and reporting tools. Vendors such as MicroStrategy and Tableau have optimized connectors for querying directly against data in Hadoop. Direct access from business intelligence tools is creating opportunities for the broader business community to work with all of the detailed data that is being stored and processed in Hadoop. Organizations can now do more with their data than when data pipelines were structured and monolithic.

Reference Data

Hadoop includes built in facilities for handling reference data. When processing transaction records, a data workflow may require reconciling against a reference data set such as a known list of accounts. Depending on the size of the data set and the type of processing, CDH includes systems for caching and joining raw data with reference data.[1]

1) For more information on using reference data in Hadoop based processing:

- Review HBase explanations on the Cloudera blog: www.cloudera.com/blog/2011/02/log-event-processing-with-hbase
- Learn about Map and Reduce side joins in Hadoop: The Definitive Guide by Tom White.

Conclusion

Hadoop is rapidly becoming a mainstay in organizations due to its flexibility, scalability and low cost of storing and processing raw data. Cloudera Enterprise — comprised of CDH, Cloudera Manager and expert support — makes Hadoop a reliable, enterprise-ready platform. With an increased focus on improving operational efficiency, leading organizations across industries are moving mission critical data processing and historical data storage to Cloudera Enterprise. This enables them to store raw data in its native format and develop and maintain complex data pipelines faster, and at significantly lower costs, than were possible using traditional systems.

About Cloudera

Cloudera, the leader in Apache Hadoop-based software and services, enables data driven enterprises to easily derive business value from all their structured and unstructured data. As the top contributor to the Apache open source community and with tens of thousands of nodes under management across customers in financial services, government, telecommunications, media, web, advertising, retail, energy, bioinformatics, pharma/healthcare, university research, oil and gas and gaming, Cloudera's depth of experience and commitment to sharing expertise are unrivaled.