Real-time Data Pipelines: Bring in More Data, Faster

Challenges by Role

Data Engineer
Rapidly adapting ecosystem of tools and technology requires additional expertise and new considerations to the company's strategy.

Data Scientist
Data discovery and application development requires access to more data in less time. Data Scientists depend on the data engineer to bring in the data they need and make it accessible to them.

Enterprise Architect
New data collection requires additional technology in order to support the growing needs of the business, and these technologies need to carry the business into the future. Data pipelines and defined processes are needed.

Operations
Meeting service level guarantees becomes impossible when data volumes become too large and the number and size of processing jobs grows, often resulting in system latency.

Introduction

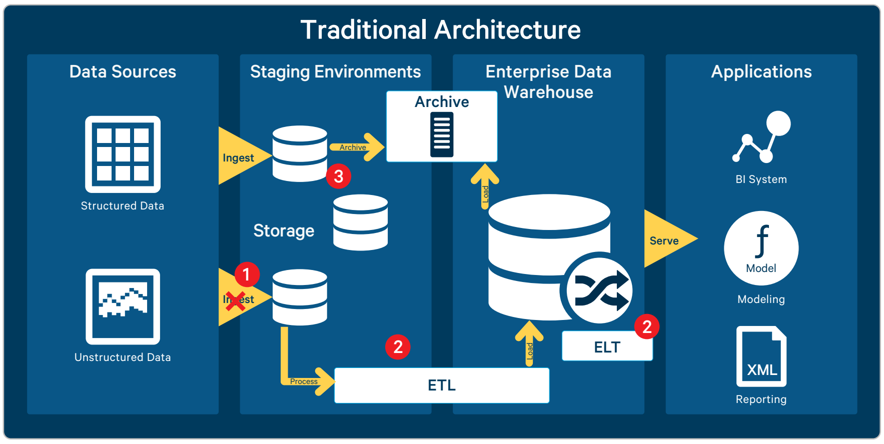
Data is fueling a world of opportunity. Increasingly, forward thinking organizations are placing their focus on data to drive internal strategy and deliver superior service capabilities to their users. As decision making moves to real-time, systems need enhanced capabilities to meet the demands of modern data collection, transformation, and storage. Many businesses have systems that can bring in data and prepare it for analysis but some data formats are still not supported. Which begs the question, how much more accurate could your business intelligence and customer analytics be if you were able to leverage more of the relevant data to your business? Or how much more accurate could your business forecasts be if you could keep data online longer?

As storage technology advances to meet the demand of keeping more data, Apache Hadoop offers users the ability to leverage streaming, interactive, and complex data types. Cloudera Enterprise is a platform that gives users the tools they need to ingest, transform, and cleanse a wide variety of data and create end-to-end data pipelines. The platform also provides integration with many popular ETL and data wrangling tools so engineers have the right tools they need to be efficient, with minimal business disruption.

Data and ETL Engineers are constantly tasked with delivering data in a presentable analytic format within a certain time commitment to the business. Often these transformation windows can slip into alarming territory when handling new data sources. These processes can only be successful if the systems, technology, and teams are in alignment to handle the rise in data collection, and any misalignment can result in missed commitments.

Challenges

Augmenting your current architecture and equipping your team to develop real-time data pipelines requires a robust toolset that can meet the demands of a wide variety of data. Unstructured data, time-series data, and streaming data was often simply left out of the equation due to the complexity of bringing it in and transforming it so that businesses can utilize it for analysis. Today's businesses need the ability to keep more data online and leverage new data as fast as they can bring it in. Many solutions that facilitate new data collection often had no ability to deal with historical data or cratered under the pressure of complex data types. The most common limitations include:



Limited Data

Traditional data processing systems are limited to the constraints of the downstream data warehouse. Data is archived or even deleted when systems reach capacity making it inaccessible to users. Many times complex data types were left out of analysis activities due to an inability to combine incoming data types with the organization's core data. As data volumes scale, it increases the load on critical data management systems, which results in diminished performance.

Poor Performance

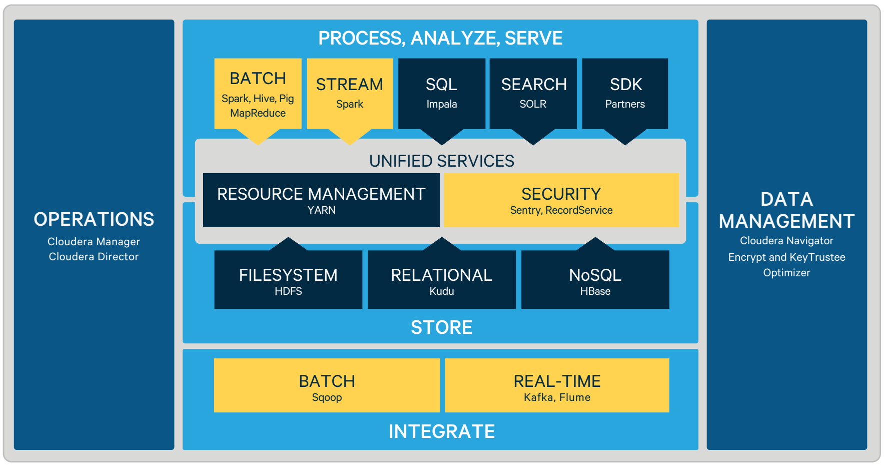
Engineers often have a service level guarantee (SLA) to provide data to the business. When processing windows encroach the service commitment, it leaves little time to ensure data quality and common standards are met. It also may result in an inability to provide requested views to the business bolstering low faith in new data projects. When we consider the demands of streaming data or time relevant data most modern solutions simply don't have the processing resources to meet the demand, especially within the SLA confines.

Operational Complexity

Matching the skills of your engineering team to the ever-evolving ecosystem of tooling and programing languages has led to organizational skills gaps and siloed data systems. New systems and integrations can produce management overhead as well as gaps in security and governance. Introducing new functionality means equipping your team with the experts to access and process that data. Once the technology is addressed, the tools the team needs to be successful must be available.

Solution

Modernization of your data ingestion will require an approach with a comprehensive toolset and the ability accelerate common third-party tooling coupled with the performance characteristics to meet the demands of bringing in more data. With Cloudera Enterprise, organizations are able to leverage a single, unified platform in order to ingest, transform, and serve a wide variety of data to power your business. Cloudera's leadership in development, training, and services around powerful new processing tools such as Apache Spark showcase a commitment to Spark's role in the future (and constant innovation) of the Hadoop ecosystem. Customer choose Cloudera because we are the best positioned to help users increase performance and reduce operational complexity.



Integrate More Data, Faster

Cloudera Enterprise allows you to collect, store, and process unlimited data to give the most comprehensive view of all your data. With included tools for ETL, data transformation, and connectors to all the leading database formats; new data sources can be adopted as the business dictates. Cloudera's platform can handle a variety of real-time formats with the capable processing technology to handle even the most demanding high-velocity use cases. Better integration of more data helps build the best analytic outputs across the business.

High Performance

With new tools that leverage modern hardware profiles like Apache Spark, companies can turn their processing timelines from hours or days into minutes. Reduced time and complexity to bring in new data sources allows engineers to focus on new functionality and identification of potential data sources. As an added bonus, reduced performance load on data warehouse systems for ELT workloads means users can focus these systems on what they're optimized to do, including querying and reporting.

Reduced Operational Complexity

Cloudera Enterprise gives you the flexibility to work with data however you need to; through interactive SQL, batch processing, or full-text search. This increases developer access to incoming data while affording direct access for common SQL and third party tooling for reporting. Cloudera has extended support for popular Python and Scala tooling to widen the adoption capabilities of engineering tools. All of this in a single technology platform with unified data, metadata, security, and governance. A single platform for you to manage and master to provide enhanced returns on new data and support evolving use cases. Cloudera also offers industry leading training and services to equip your teams to meet the need faster.

Summary

Stop settling for limited analytic capabilities, poor access to data, and expensive system scaling. Afford your team the right tools and increased performance they need to be successful with new data types. Your business needs more data in order to better understand your customers, offerings, and threats to your business. Your business strategy dictates the need for data to drive critical decisions. Cloudera Enterprise is engineered to accelerate data ingestion and offers users the right tool for their workload. Cloudera's leadership in Apache Spark is helping make real-time pipelines a reality while enterprise management and administration tools are giving users data visibility and end-to-end security so they bring in more data without introducing new risk. Cloudera Enterprise makes modern data processing fast, easy, and secure.

What's Next?

- Data Engineering Website
- Get Certified

cloudera®

About Cloudera

Cloudera delivers the modern platform for data management and analytics. The world's leading organizations trust Cloudera to help solve their most challenging business problems with Cloudera Enterprise, the fastest, easiest, and most secure data platform built on Apache Hadoop. Our customers can efficiently capture, store, process, and analyze vast amounts of data, empowering them to use advanced analytics to drive business decisions quickly, flexibly, and at lower cost than has been possible before. To ensure our customers are successful, we offer comprehensive support, training, and professional services. Learn more at cloudera.com.

cloudera.com

1-888-789-1488 or 1-650-362-0488
Cloudera, Inc. 1001 Page Mill Road, Palo Alto, CA 94304, USA