cloudera

Hive, Impala, and Spark, Oh My: SQL-on-Hadoop in Cloudera 5.5

Justin Erickson | Director of Product Management | Cloudera

Agenda

- History of SQL-on-Hadoop technologies
- Picking the right tool for the job
- What's new with Cloudera 5.5
- Real-world use cases
- Future of SQL-on-Hadoop

cloudera

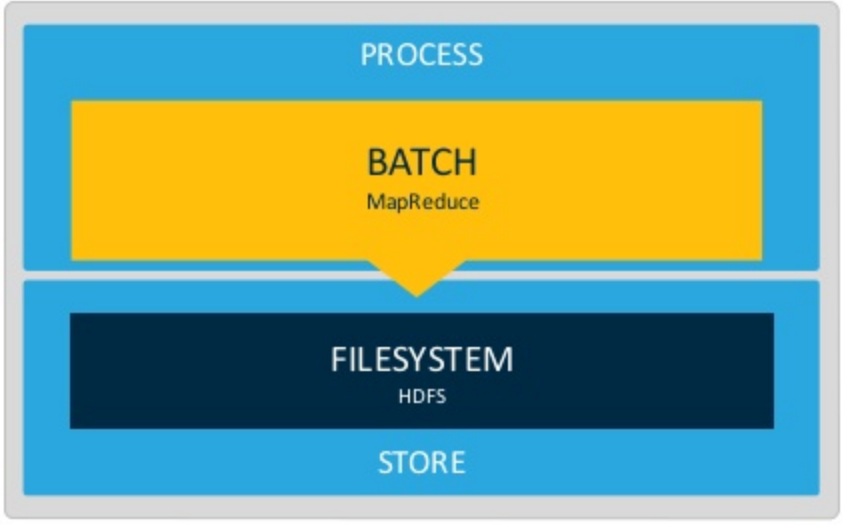
MapReduce: The Early Years

The original processing engine for Hadoop

- Process any type of data in any format
- Scale infinitely for multiple, large jobs
- Pioneer of bringing compute to data

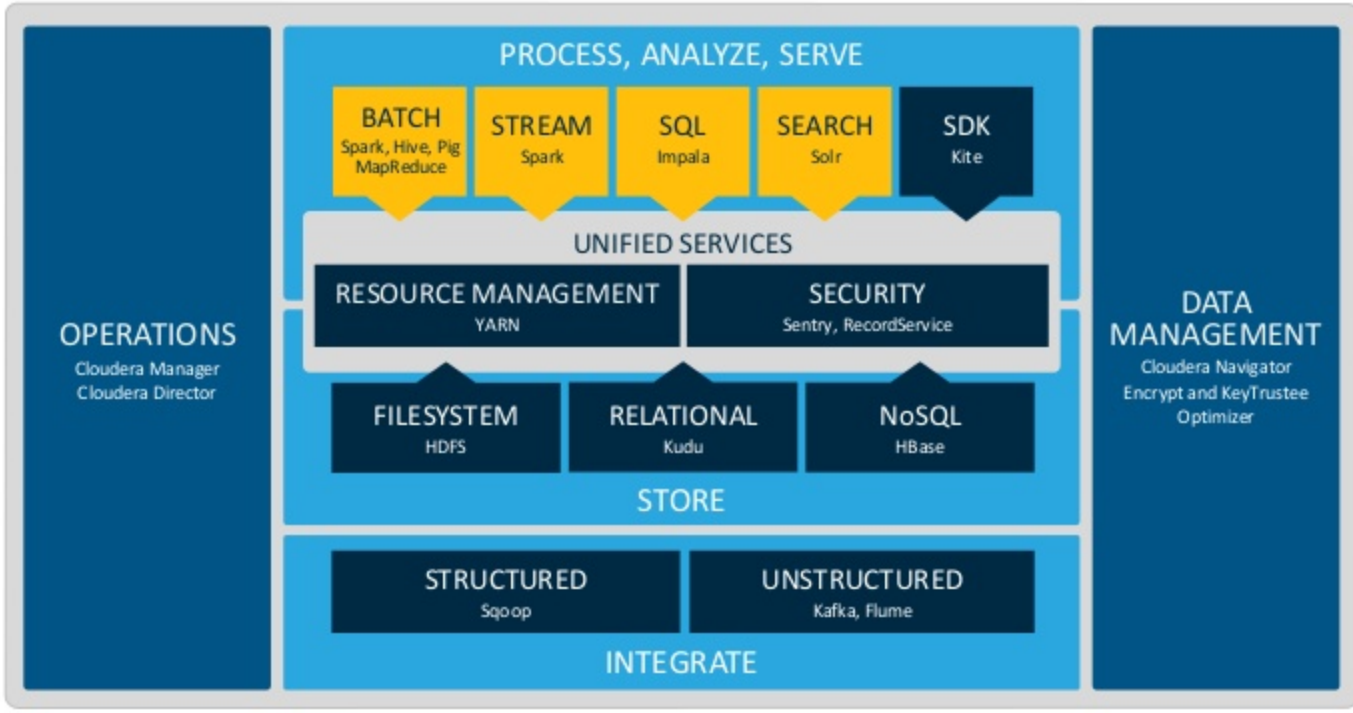
But…

- Difficult to program
- Slow processing
- Limited expressivity



cloudera

One Platform, Many Workloads



Batch, Interactive, and Real-Time.

Leading performance and usability in one platform.

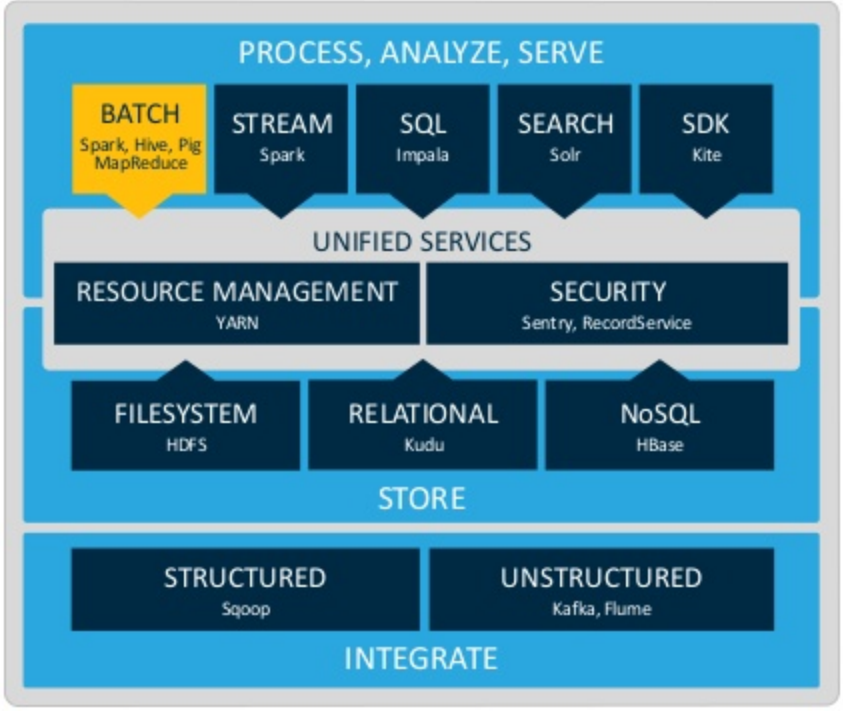
- End-to-end analytic workflows
- Access more data
- Work with data in new ways
- Enable new users

cloudera

The Need for SQL for Batch Processing

Apache Hive

- Eases development on MapReduce with familiar SQL
- Built for long-running ETL, data preparation, and batch processing
- Shared data structures across Hadoop tools

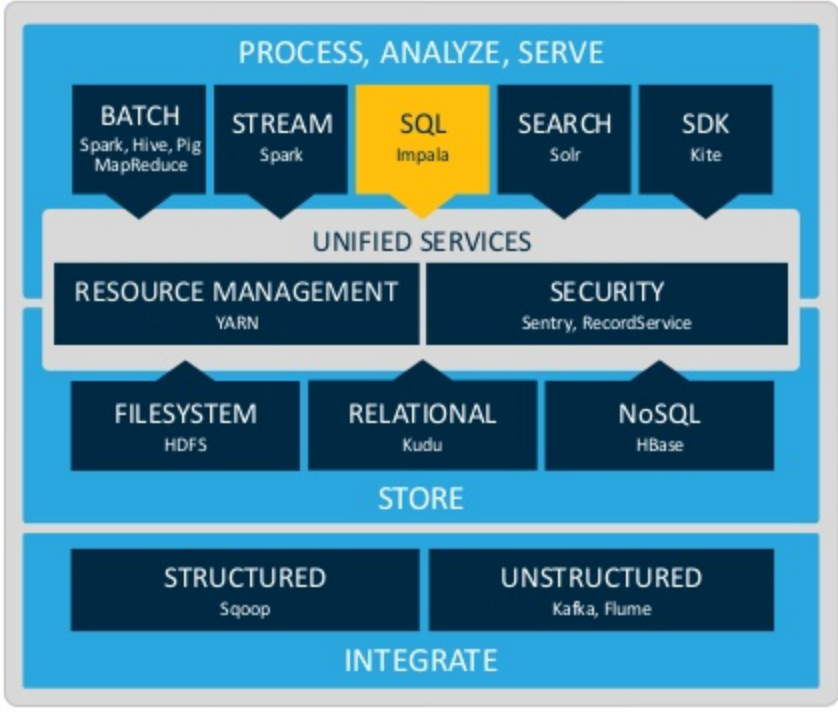


cloudera

The Need for Interactive SQL for BI

Apache Impala (incubating)

- Low latency for interactive performance
- Built for multi-user workloads
- Compatible with SQL and leading BI partner tools

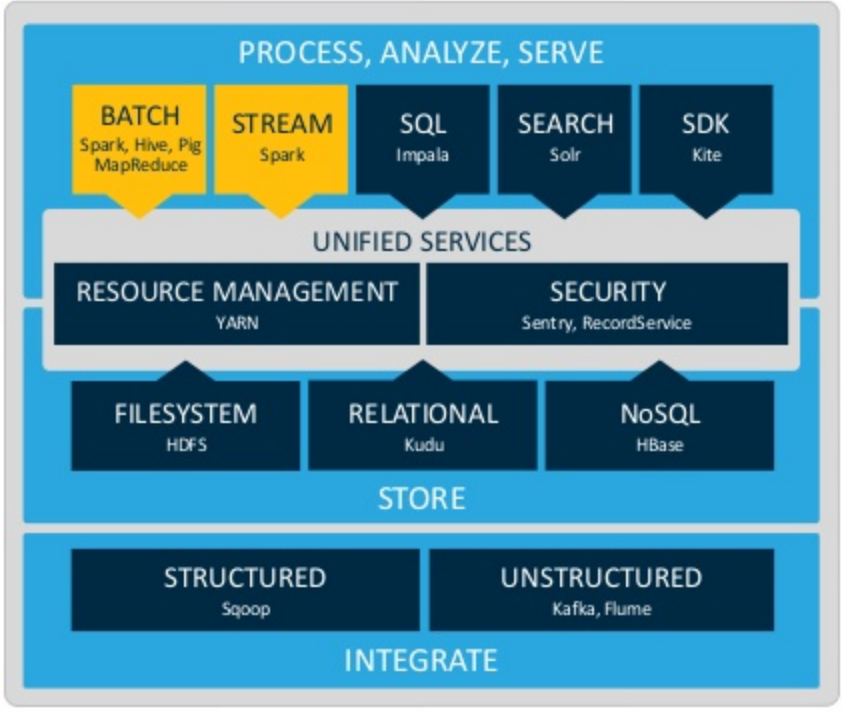


cloudera

The Need for Flexible Data Processing

Apache Spark (and Spark SQL)

- Easy development
- Flexible, extensible API across multiple workload types
- In-memory batch and stream processing performance boost



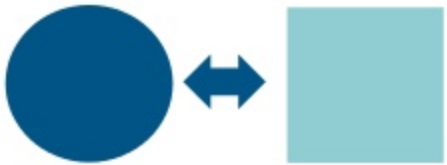
cloudera

Focus on Open Source Standards

Open source does not guarantee a future-proof investment

Long-Term Architecture

Only open standards get continuing, long-term investment from across the ecosystem.



Avoidance of Lock-in

Open standards have multi-vendor support, giving customers choices and preventing lock-in.

Ecosystem Compatibility

Open standards attract more third-party connectors/certifications due to broad adoption.

Choosing the Right SQL Engine

Know Your Audience, Know Your Use Case

Batch Processing

BI and SQL Analytics

Procedural Development

SQL-on-Hadoop in Cloudera 5.5

	Apache Hive	Apache Impala (incubating)	Apache Spark SQL
Audience	ETL Developers	Business Analysts	Data Engineers & Data Scientists
Strengths	• Built for very long-running ETL, data preparation, or batch processing • Supports custom file formats • Handles massive ETL sorts with joins	• Scales to high-concurrency • Supports high-performance interactive SQL • Compatible with BI tools & skills • Hadoop integration & usability	• Easily embed SQL into Java, Scala, or Python applications • Simple language for common operations • Seamlessly mix SQL and Spark code within a single application
New Features	• Hive in the cloud (S3) • Hive-on-Spark beta • Governance & Lineage	• Nested data types • Column-level security • Integration with Kudu (beta)	• Support for Spark SQL & DataFrames • Hive integration • Automatic performance optimizations

cloudera

SQL-on-Hadoop Benchmark

Impala, Spark SQL, Hive-on-Tez

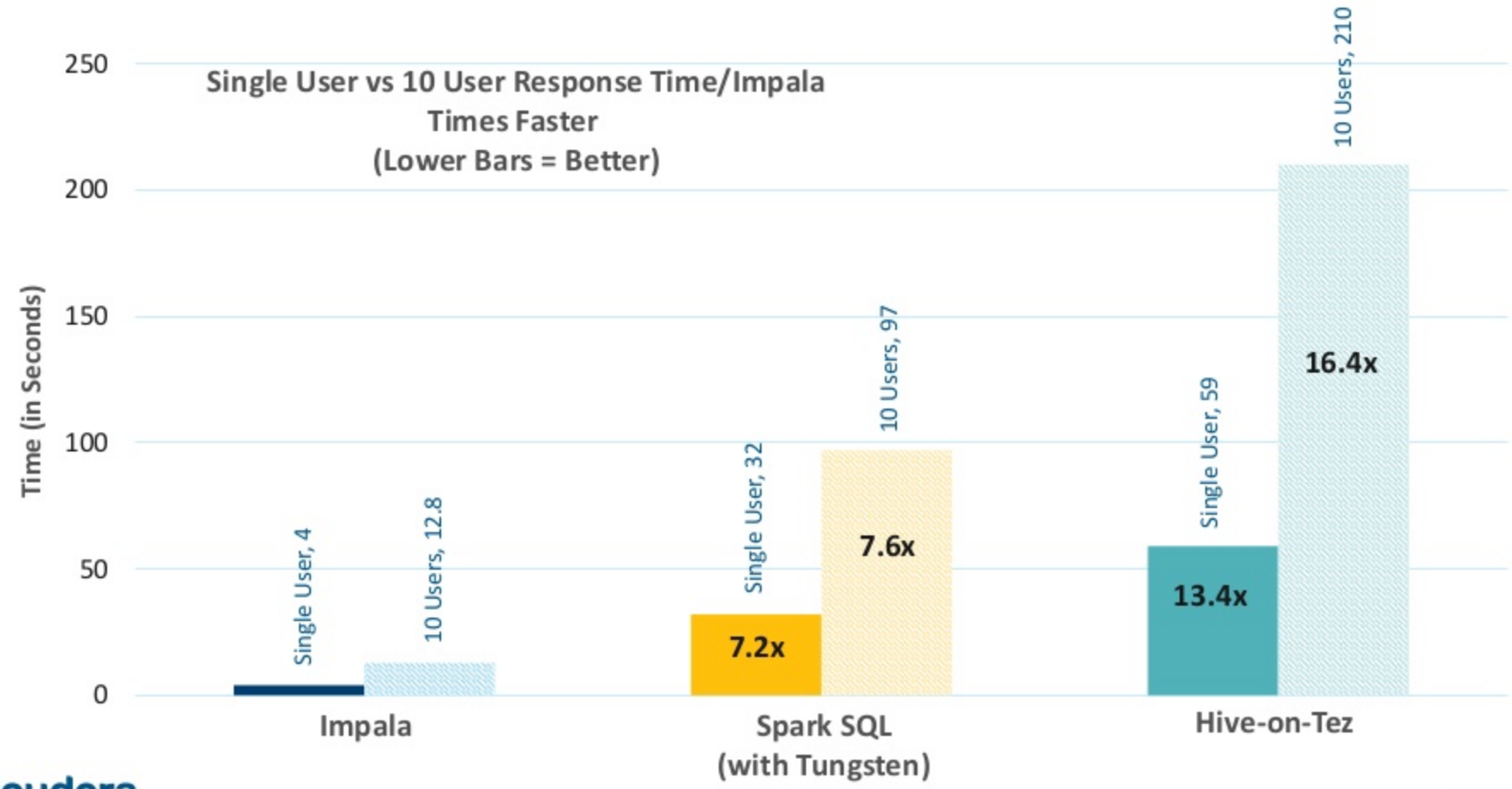
Versions:

- Impala 2.3
- Hive 2.0 on Tez 0.5.2 (aka “Stinger”)
- Spark SQL 1.5 with Tungsten

- Benchmark Details
  - Based on industry standards (TPC)
  - Repeatable
  - Methodical testing with multiple runs on same hardware
  - Help competing software do well
    - Run on optimal file formats for each
    - Tune query engines appropriately

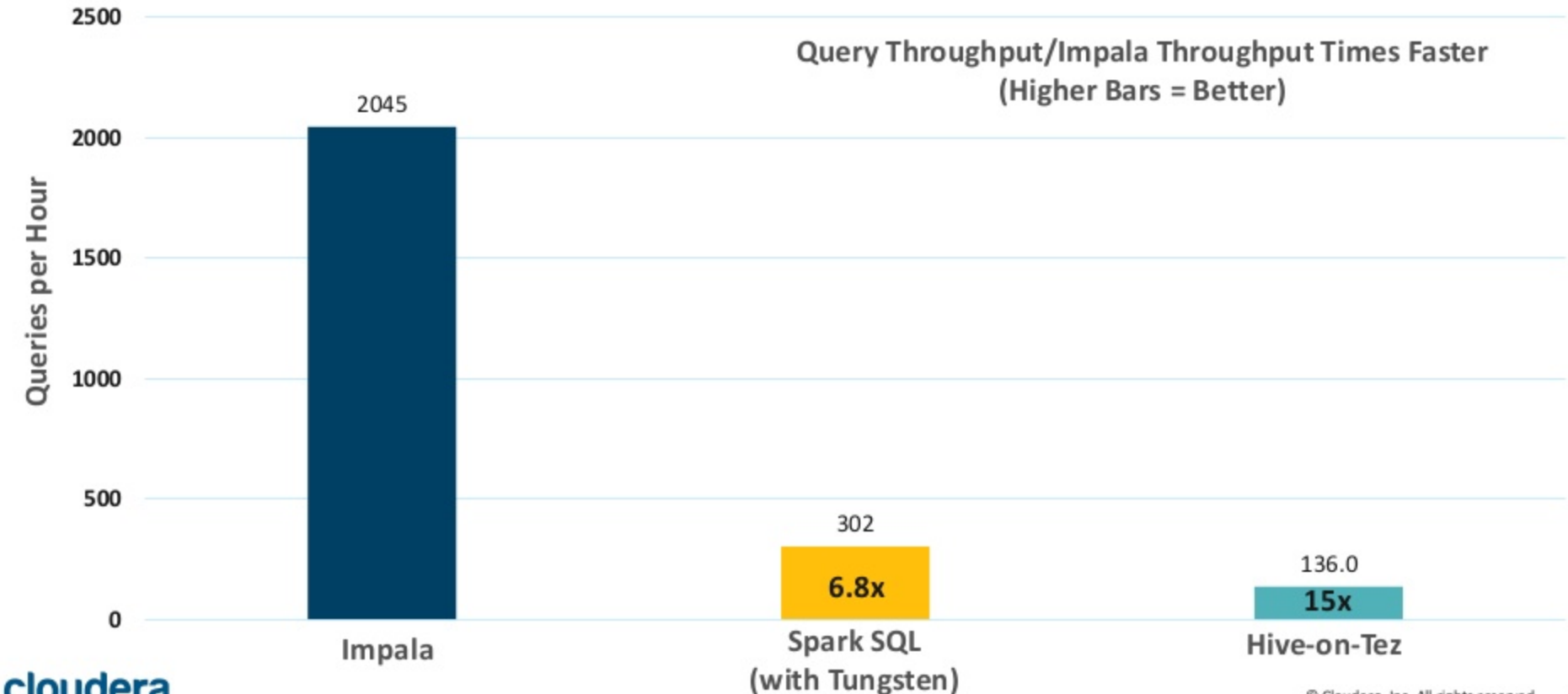
cloudera

Impala Multi-User Performance Over 7x Faster



cloudera

Impala Enables Nearly 7x Throughput



cloudera

Performance Benchmark Takeaways

- **Impala unlocks BI usage directly on Hadoop**
  - Meets BI **low-latency** and **multi-user** requirements
  - Advantage expands for single-user vs just 10 users
- **Hive is designed (and still great) for batch processing**
  - Most Impala customers use Hive for data preparation
  - Hive is the most commonly used ETL framework
- **Spark SQL enables easier Spark application development**
  - Enables mixed procedural Spark (Java/Scala) and SQL job development
- **Mid-term trends will further favor Impala's design approach for latency and concurrency**
  - More data sets **move to memory** (HDFS caching, in-memory joins, Intel joint roadmap)
  - **CPU efficiency** will increase in importance
  - Native code enables **easy optimizations for CPU** instruction sets
  - **Intel joint roadmap** support these opportunities

cloudera

Use Cases

cloudera

PREMIER

PROBLEM

Needed to efficiently collect, process, and analyze data from growing hospital network

- EDW couldn't meet scale and unstructured data demands
- Processing too slow for actionable decisions
- Limited, time consuming supply chain matching

SOLUTION

Integrated 1000s of hospital systems through unified enterprise data hub

- Ingest and process **45% more spend data**
- **Faster analytics on $41B** through end-user healthcare spend dashboard
- Unprecedented **matching of 98%** of supply chain data
- **Better TCO** through unification and licensing costs for **new opportunities**

cloudera

epsilon

PROBLEM

Clients had limited insights to thousands of marketing campaigns across channels

- Clients want real-time campaign updates with 3-sec SLA
- Existing system couldn't meet scaling or data type demands
- Limited self-service BI

SOLUTION

Built next-generation digital marketing platform for 360-degree customer view

- **Improved query performance** from minutes to seconds to meet SLAs
- **Enhanced modeling** with combined online and offline data
- Real-time optimizations through **interactive, self-service access**

cloudera

Cox Automotive™

PROBLEM

Couldn't support data integration across 20+ brands

- Existing systems couldn't scale for data consolidation
- Siloed access based on workload
- No real-time data ingestion or access

SOLUTION

Brought all data directly to the business to lower costs and open up new use cases fast

- **Reduced TCO by 50%** by consolidating over 1PB of data, adding 200M rows daily
- Enabled **real-time vs hourly** updates on ad performance
- **Optimized inventory management** through data matching and consolidation

cloudera

Impala Roadmap

2H 2015

- SQL Support & Usability
  - Nested structures
  - Kudu updates (beta)
- Management & Security
  - Record reader service (beta)
  - Finer-grained security (Sentry)
- Integration
  - Isilon support
  - Python interface (Ibis)
- Performance & Scale
  - Improved predictability under concurrency

1H 2016

- Performance & Scale
  - Continued scalability and concurrency
  - Initial perf/scale improvements
- Management & Security
  - Improved admission control
  - Resource utilization and showback
- SQL Support & Usability
  - Dynamic partitioning
  - Improved timestamp compatibility

2016

- Performance & Scale
  - >20x performance
  - Multi-threaded joins/aggregations
  - Continued scale work
- Management & Security
  - Improved YARN integration
  - Automated metadata
- Integration
  - S3 support
- SQL Support & Usability
  - Nested types with Avro
  - Date type
  - Added SQL extensions

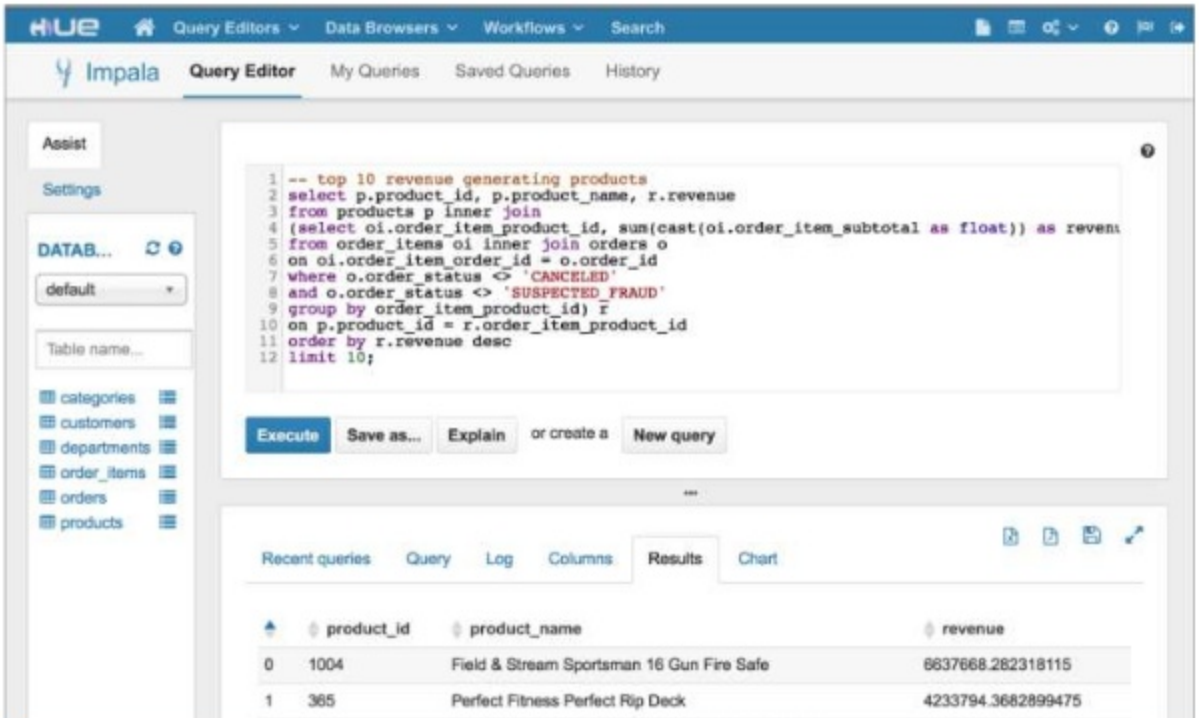
cloudera

Download Cloudera 5.5

cloudera.com/downloads

cloudera

Try It With Cloudera Live



Featuring tutorials on:

cloudera®

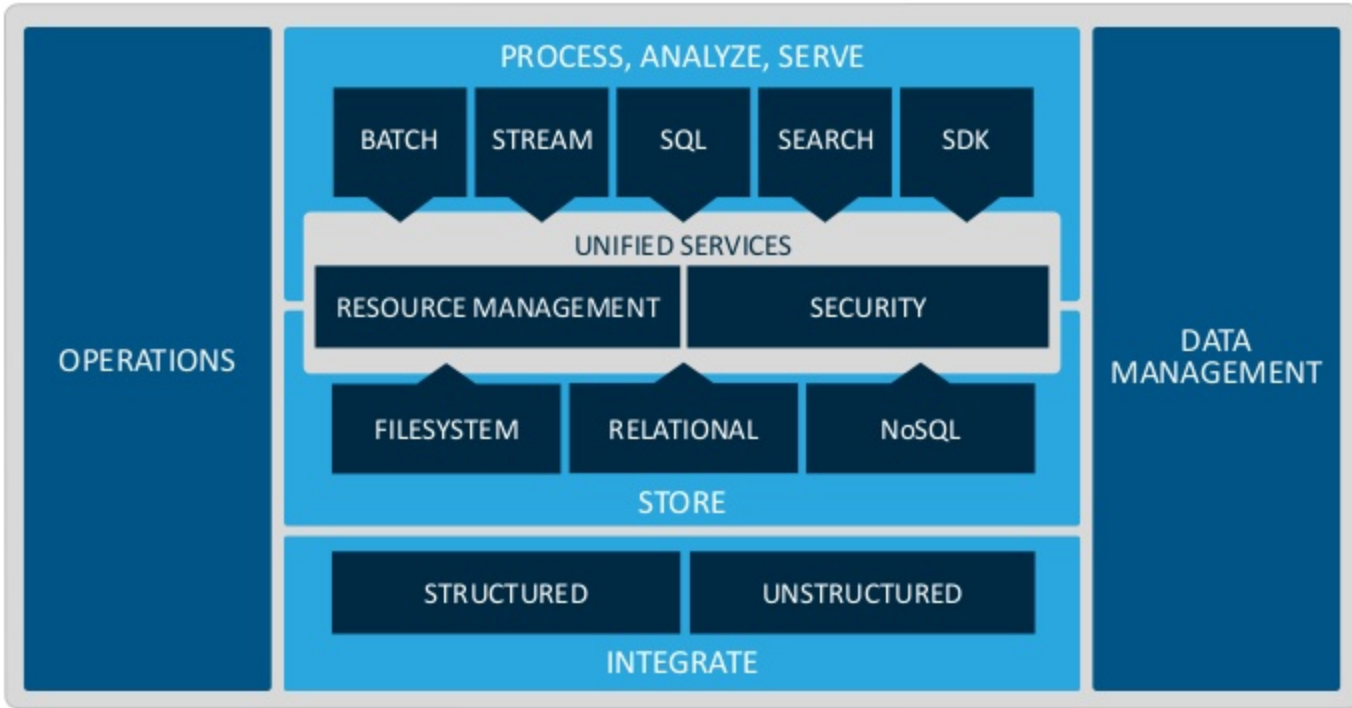
tableau SOFTWARE

ZOOMDATA®

cloudera.com/live

Cloudera Enterprise

Making Hadoop Fast, Easy, and Secure



A new kind of data platform:

- One place for unlimited data
- Unified, multi-framework data access

Cloudera makes it:

- **Fast** for business
- **Easy** to manage
- **Secure** without compromise

cloudera

cloudera

Thank You!