Cloudera Impala

Justin Erickson | Senior Product Manager

May 2013

Agenda

- Why Impala?
- Architectural Overview
- Real-World Use Cases
- Alternative Approaches
- The Platform for Big Data

Why Hadoop?

- **Scalability**
  - Simply scales just by adding nodes
  - Local processing to avoid network bottlenecks

- **Flexibility**
  - All kinds of data (blobs, documents, records, etc)
  - In all forms (structured, semi-structured, unstructured)
  - Store anything ***then later*** analyze what you need

- **Efficiency**
  - Cost efficiency (<$1k/TB) on commodity hardware
  - Unified storage, metadata, security (no duplication or synchronization)

What’s Impala?

- **Interactive SQL**
  - Typically 5-65x faster than Hive (observed up to 100x faster)
  - Responses in seconds instead of minutes (sometimes sub-second)
- **Nearly ANSI-92 standard SQL queries with Hive SQL**
  - Compatible SQL interface for existing Hadoop/CDH applications
  - Based on industry standard SQL
- **Natively on Hadoop/HBase storage and metadata**
  - Flexibility, scale, and cost advantages of Hadoop
  - No duplication/synchronization of data and metadata
  - Local processing to avoid network bottlenecks
- **Separate runtime from MapReduce**
  - MapReduce is designed and great for batch
  - Impala is purpose-built for low-latency SQL queries on Hadoop

Benefits of Impala

More & Faster Value from “Big Data”

- BI tools impractical on Hadoop before Impala
- Move from 10s of Hadoop users per cluster to 100s of SQL users
- No delays from data migration

Flexibility

- Query across existing data
- Select best-fit file formats (Parquet, Avro, etc.)
- Run multiple frameworks on the same data at the same time

Cost Efficiency

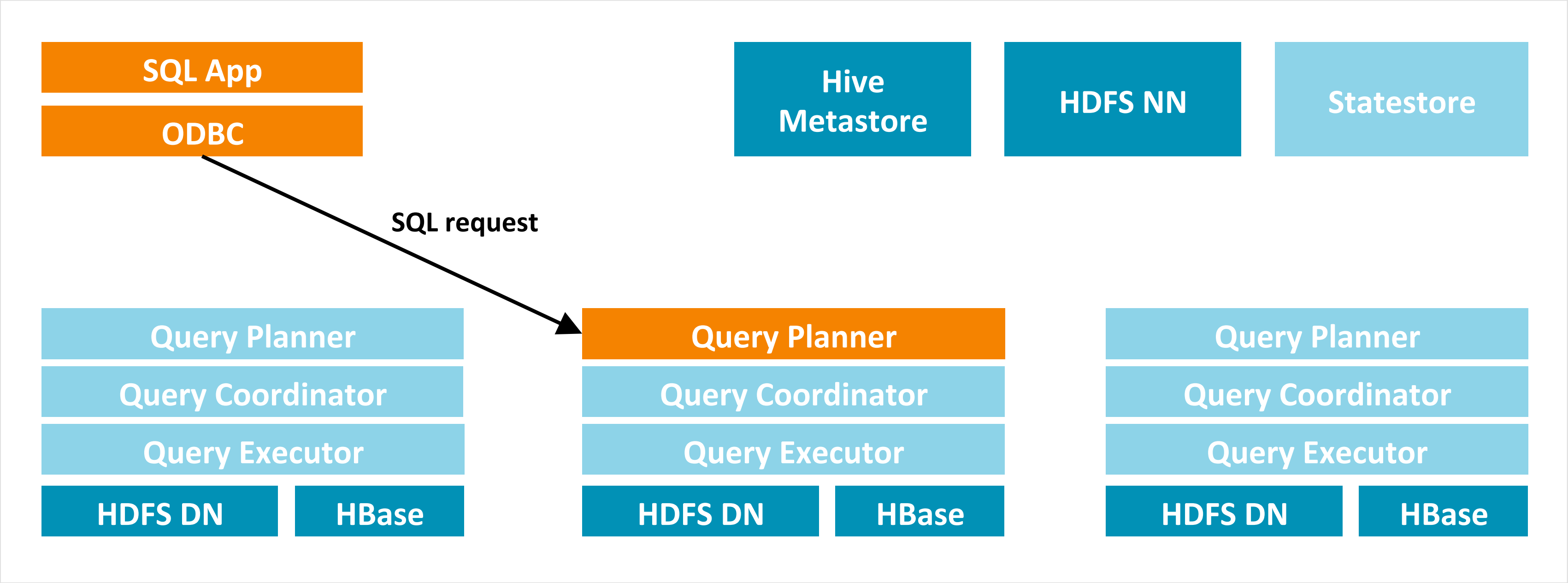
- Reduce movement, duplicate storage & compute
- 10% to 1% the cost of analytic DBMS

Full Fidelity Analysis

- No loss from aggregations or fixed schemas

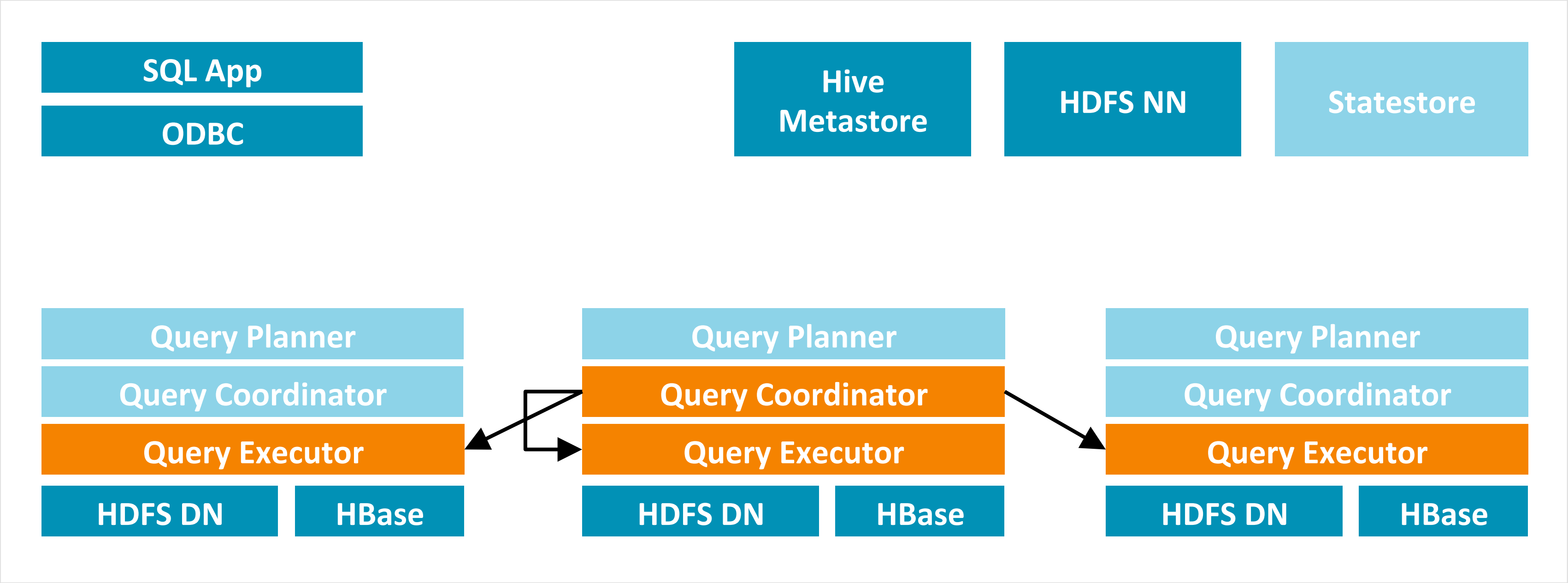
Impala Query Execution

1) Request arrives via ODBC/JDBC/Beeswax/Shell



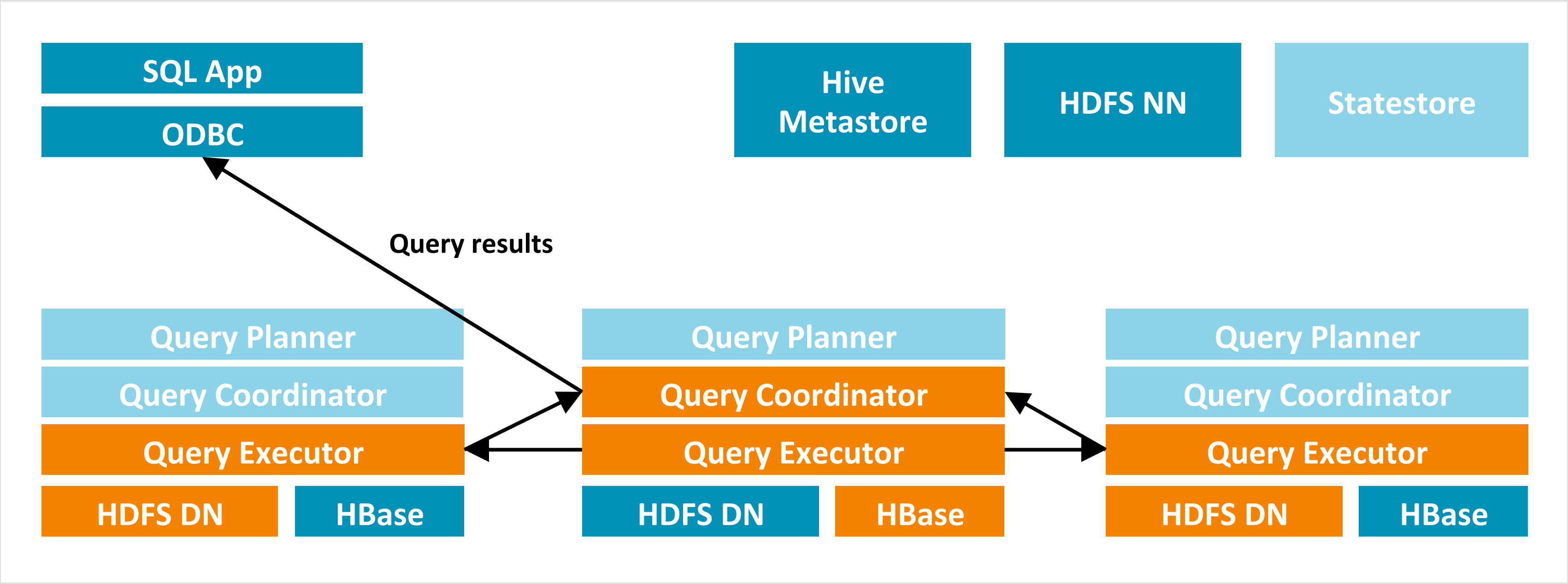
Impala Query Execution

2) Planner turns request into collections of plan fragments
3) Coordinator initiates execution on impalad(s) local to data



Impala Query Execution

4) Intermediate results are streamed between impalad(s)
5) Query results are streamed back to client



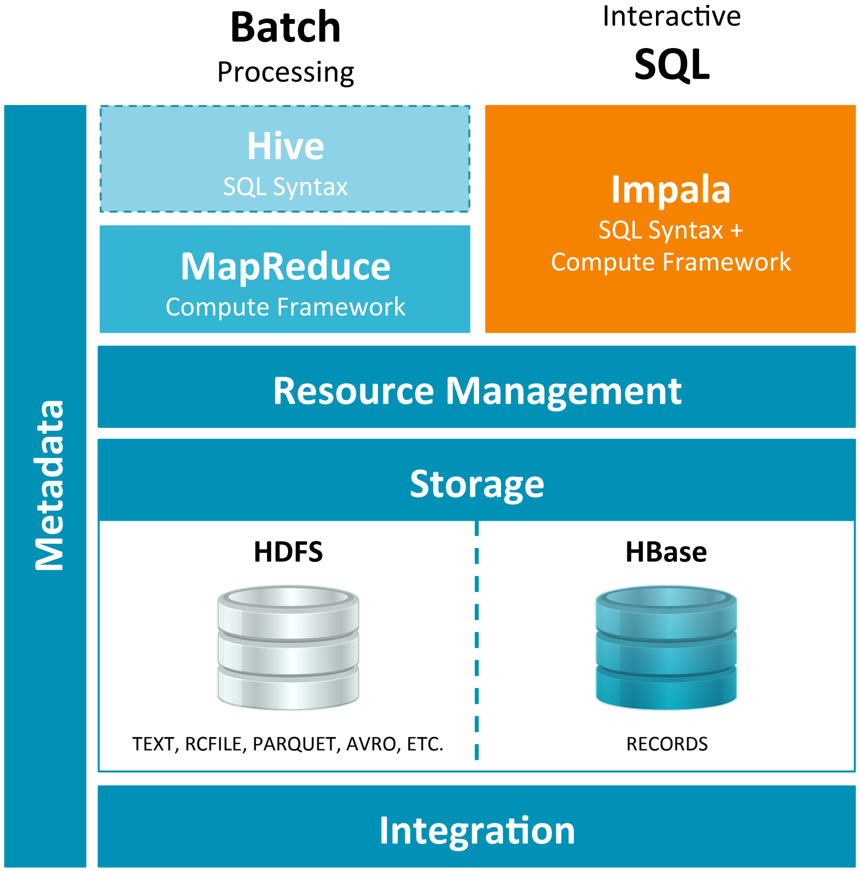
Impala and Hive

Shares Everything Client-Facing

- Metadata (table definitions)
- ODBC/JDBC drivers
- SQL syntax (Hive SQL)
- Flexible file formats
- Machine pool
- Hue GUI

But Built for Different Purposes

- **Hive:** runs on MapReduce and ideal for batch processing
- **Impala:** native MPP query engine ideal for interactive SQL



Impala Use Cases

Cost-effective, ad hoc query environment that offloads the data warehouse for:

- Interactive BI/analytics on more data
- Asking new questions
- Query-able archive w/ full fidelity
- Data processing with tight SLAs

Global Financial Services Company

Saving 90% on incremental EDW spend & improving performance by 5x

Offload data warehouse for query-able archive

Store decades of data cost-effectively

Process & analyze on the same system

Improve capabilities through interactive query on more data

Six3 Systems

Boosting performance by 20X for mission-critical, real-time cyber security

Analyze unstructured data with flexibility & real-time response

Integrate with existing desktop & BI tools

Deploy in minutes with Cloudera Manager

Expedia

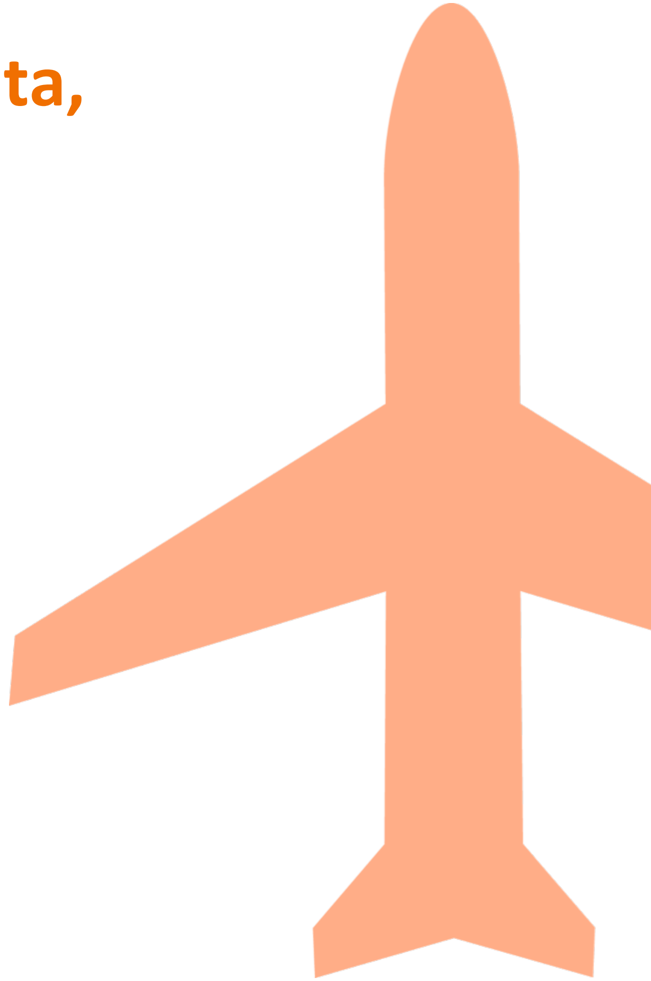
Implementing self-service BI on big data, reducing data latency by 50%

Offload data warehouse for archiving, ETL & analytics

Unify IT environment

Continuously ingest & analyze at scale

Drive greater usability & adoption of big data stack



Our Design Strategy

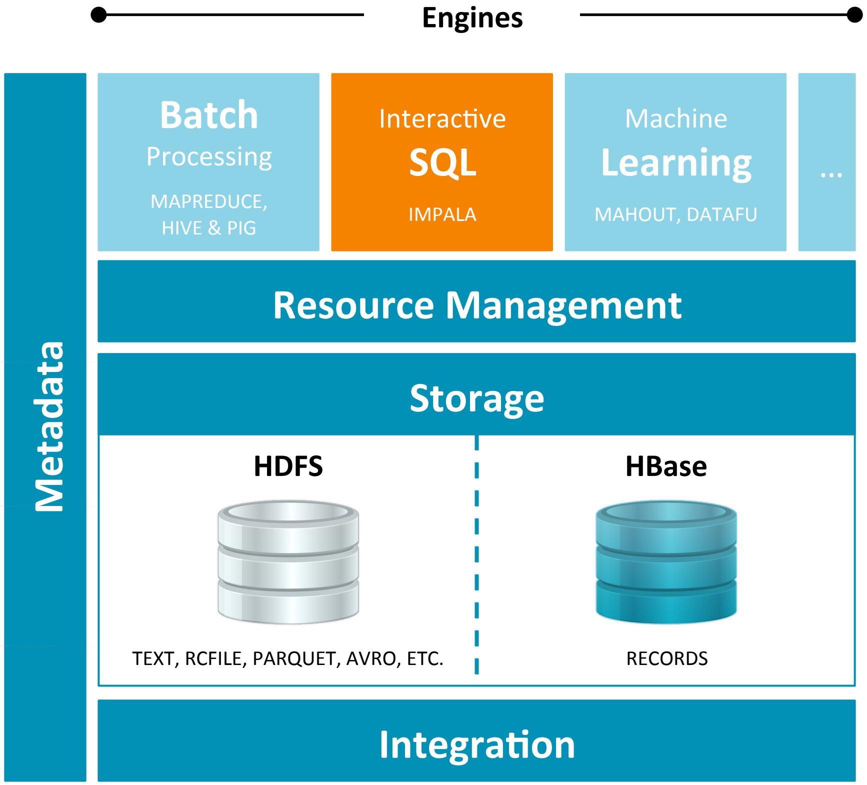
An Integrated Part of the Hadoop System

One pool of data

One metadata model

One security framework

One set of system resources



Not All SQL on Hadoop is Created Equal

Batch MapReduce

Make MapReduce faster

Slow, still batch

Remote Query

Pull data from HDFS over the network to the DW compute layer

Slow, expensive

Siloed DBMS

Load data into a proprietary database file

Rigid, siloed data, slow ETL

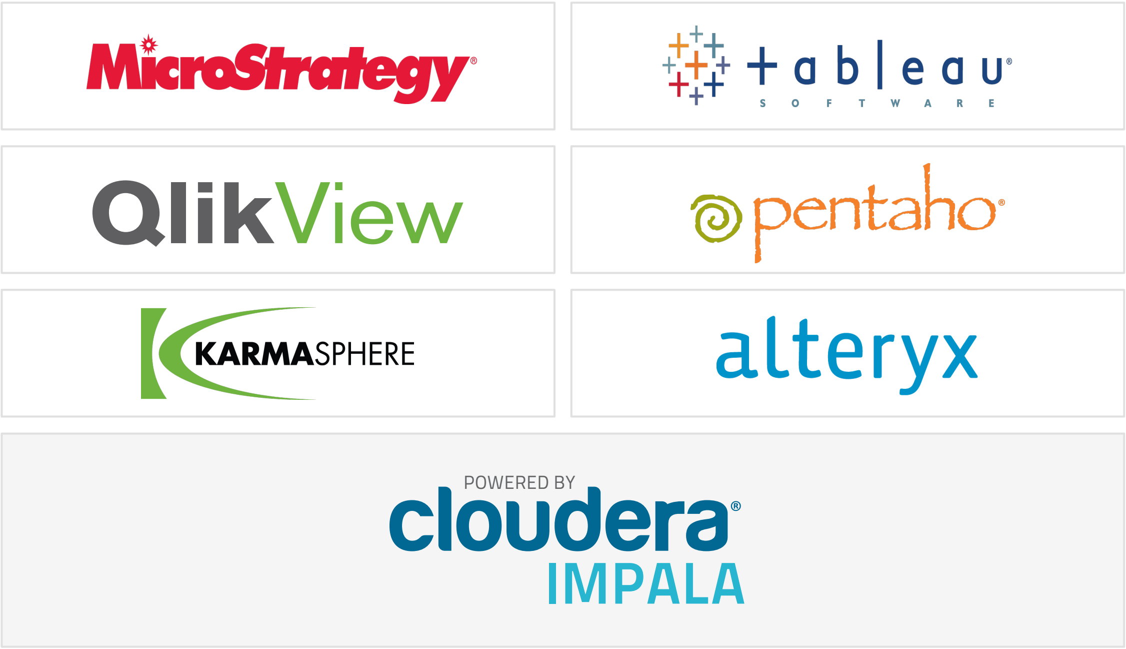
Impala

Native MPP query engine that's integrated into Hadoop

Fast, flexible, cost-effective

The Impala Advantage

BI Partners:
Building on the
Enterprise Standard



It's Not Just About SQL on Hadoop

The Platform for Big Data

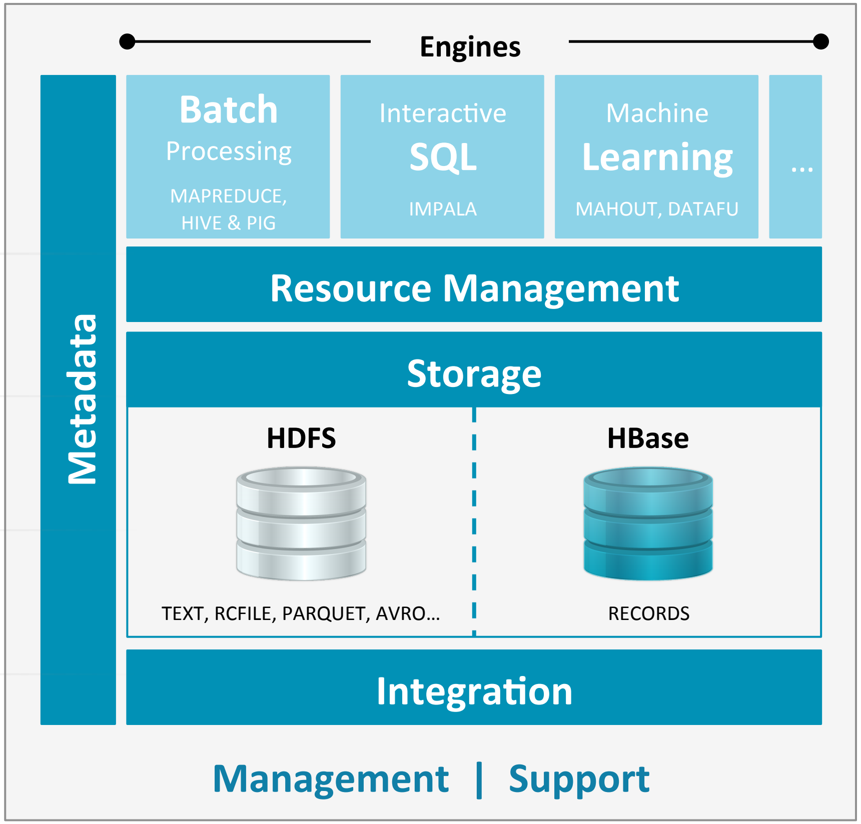
Single platform for processing & analytics

Scales to '000s of servers

No upfront schema

10% the cost per TB

Open source platform



cloudera®
Ask Bigger Questions