

# Extending Your Data Infrastructure with Hadoop

Mark Grover

Software Engineer, Cloudera

@mark\_grover

November 16th, 2013



©2013 Cloudera, Inc. All Rights Reserved.

# What I'll be Talking About

---

- Big data challenges with current data integration approaches.
- How is Hadoop being leveraged with existing data infrastructures?
- Hadoop integration – the big picture.

# What is Apache Hadoop?

**Apache Hadoop** is an open source platform for data storage and processing that is...

- ✓ Scalable
- ✓ Fault tolerant
- ✓ Distributed

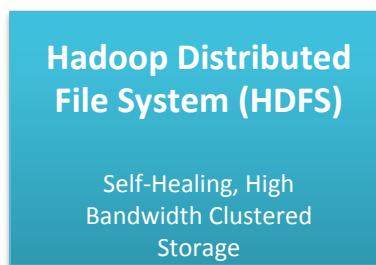
## Has the Flexibility to Store and Mine Any Type of Data

- Ask questions across structured and unstructured data that were previously impossible to ask or solve
- Not bound by a single schema

## Excels at Processing Complex Data

- Scale-out architecture divides workloads across multiple nodes
- Flexible file system eliminates ETL bottlenecks

## CORE HADOOP SYSTEM COMPONENTS



**MapReduce**  
Distributed Computing Framework

## Scales Economically

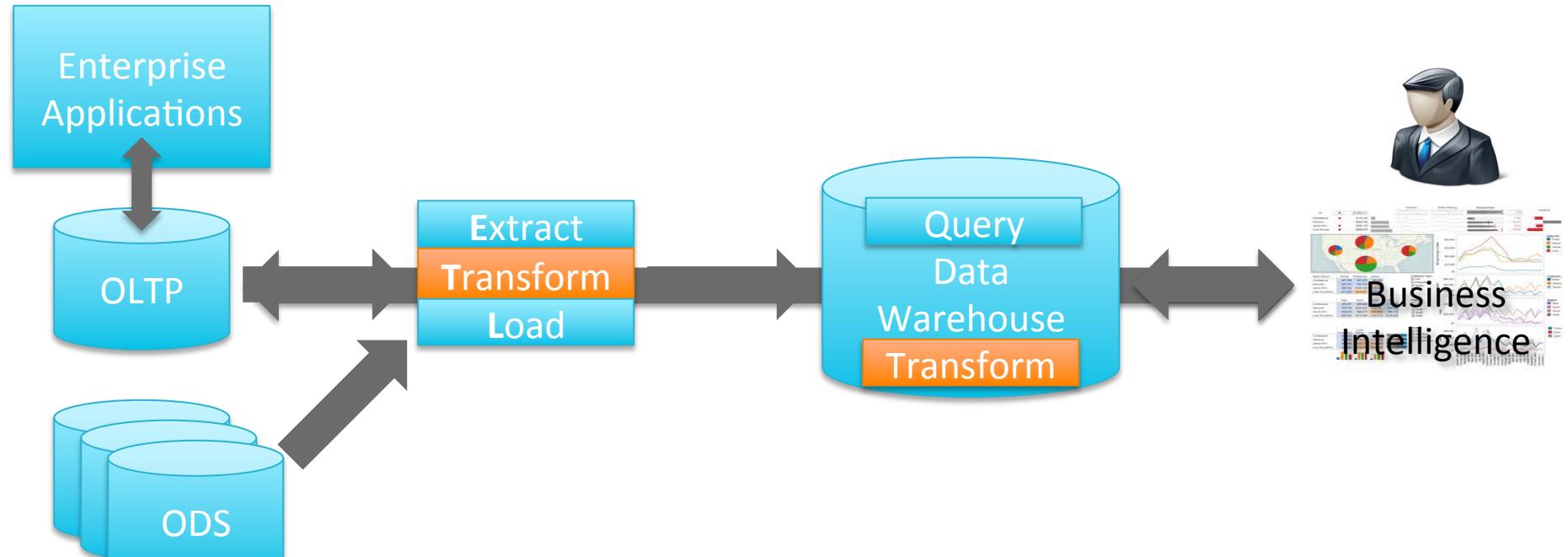
- Can be deployed on commodity hardware
- Open source platform guards against vendor lock

# Current Challenges

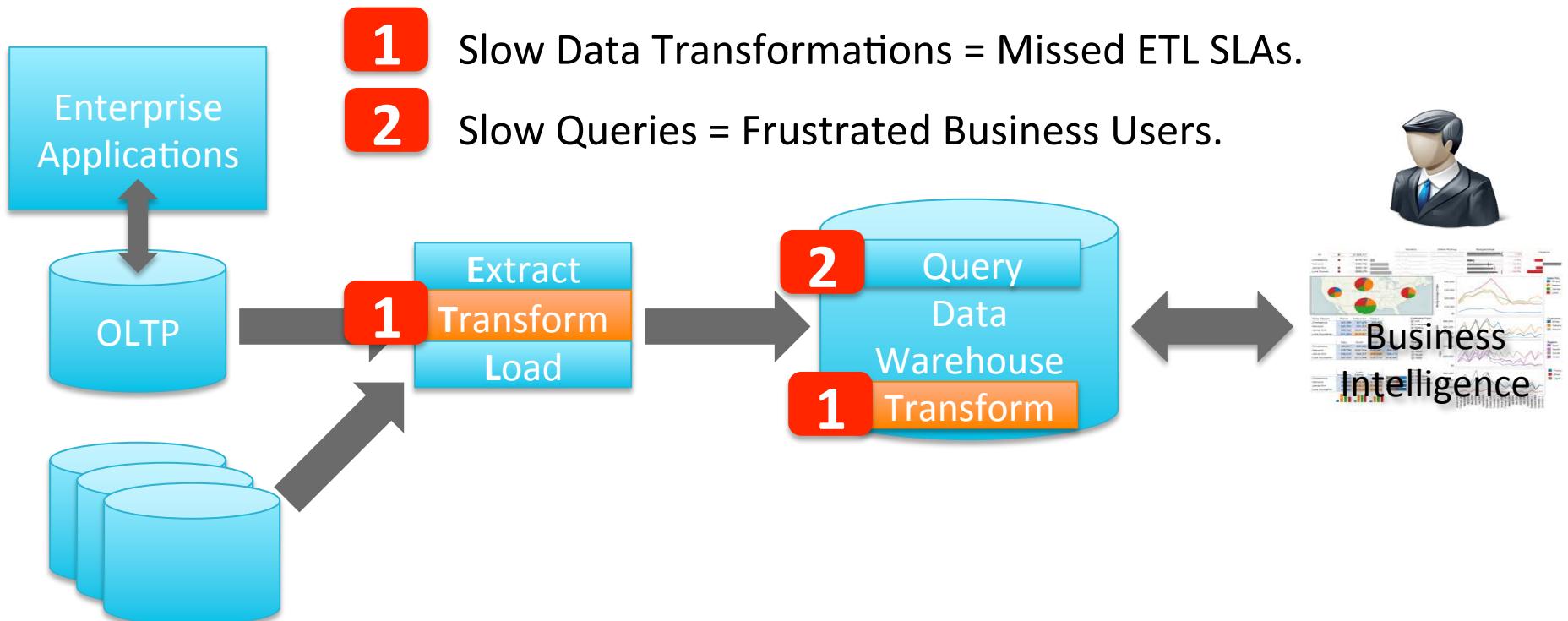
---

Limitations of Existing Data Management Systems

# The Transforming of Transformation



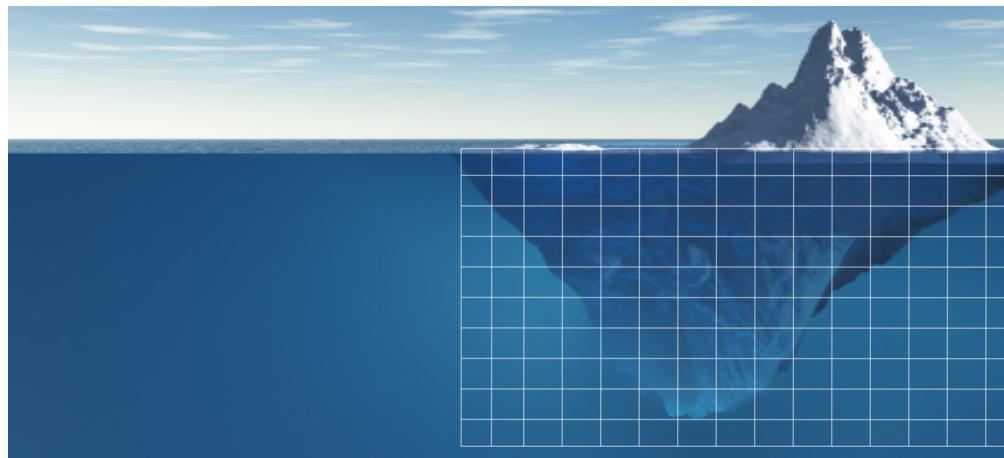
# Volume, Velocity, Variety Cause Capacity Problems



# Economics: Return on Byte

Return on Byte (ROB) =

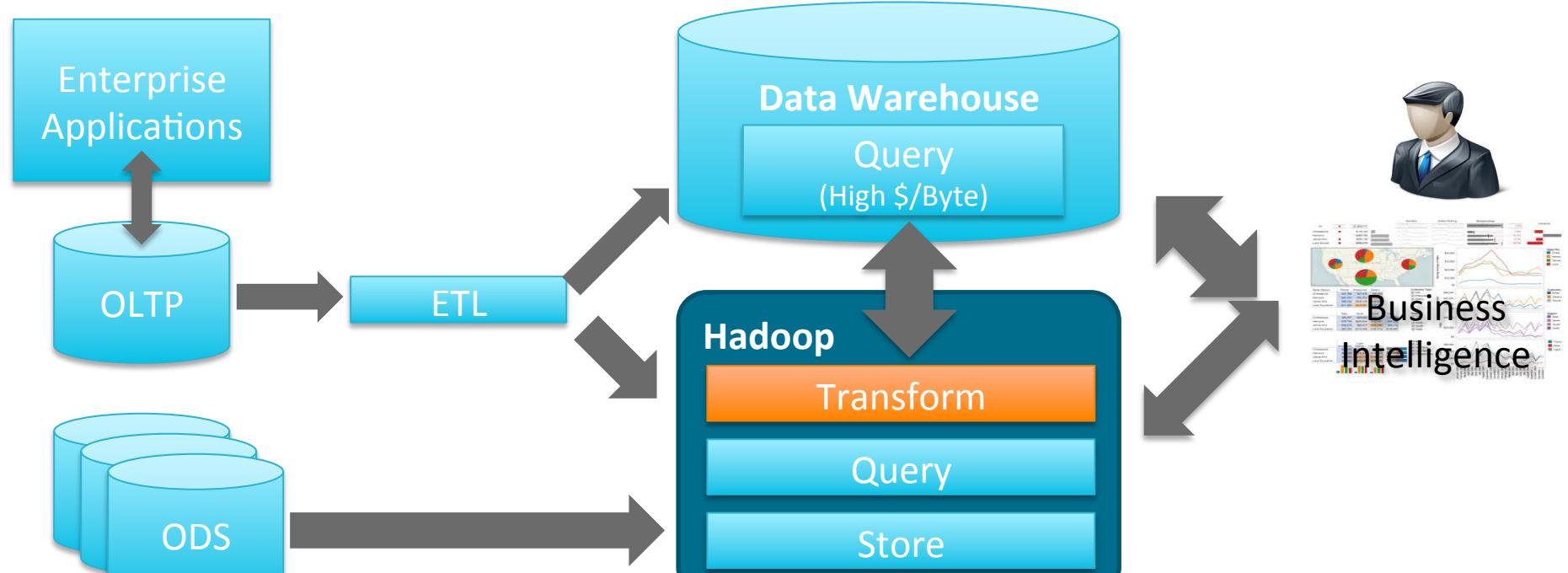
$$\frac{\text{Value of Data}}{\text{Cost of Storing Data}}$$



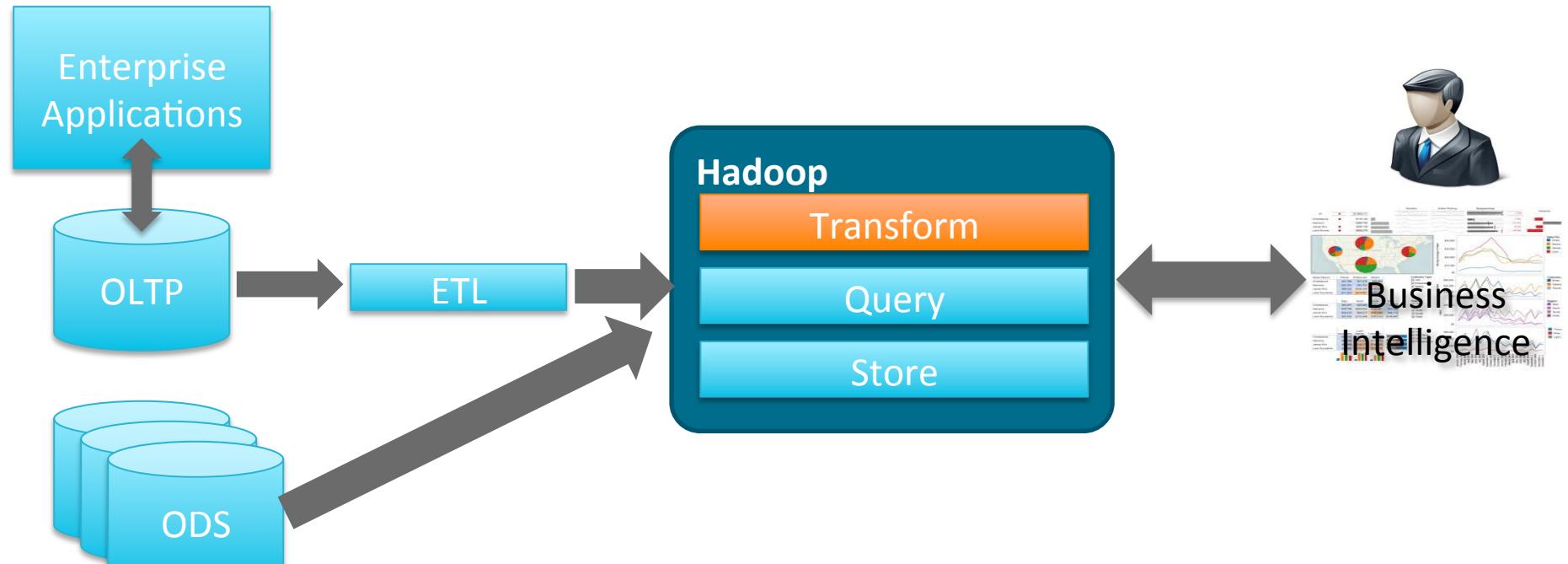
**High ROB**

**Low ROB**  
(but still a ton  
of aggregate  
value)

# Data Warehouse Optimization



# Data Warehouse Optimization



# The Key Benefit: Agility/Flexibility

## Schema-on-Write (RDBMS):

- **Prescriptive Data Modeling:**
  - Create static DB schema
  - Transform data into RDBMS
  - Query data in RDBMS format
- New columns must be added explicitly before new data can propagate into the system.
- **Good for Known Unknowns (Repetition)**

## Schema-on-Read (Hadoop):

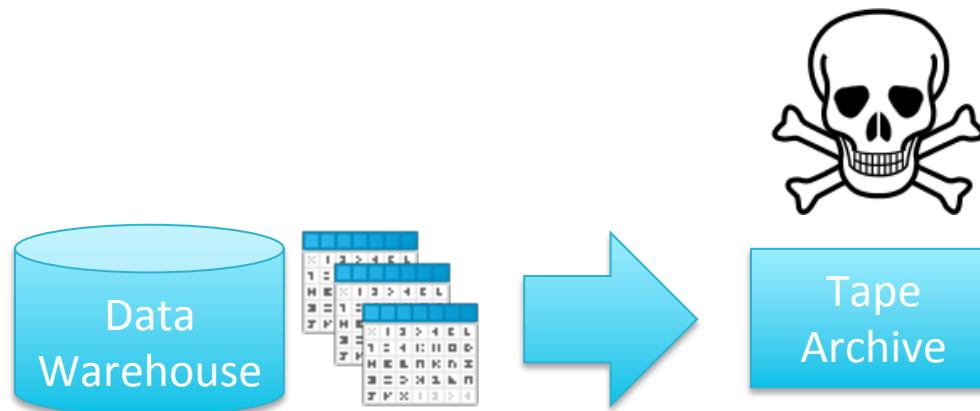
- **Descriptive Data Modeling:**
  - Copy data in its native format
  - Create schema + parser
  - Query Data in its native format
- New data can start flowing any time and will appear retroactively once the schema/parser properly describes it.
- **Good for Unknown Unknowns (Exploration)**

# Not Just Transformation

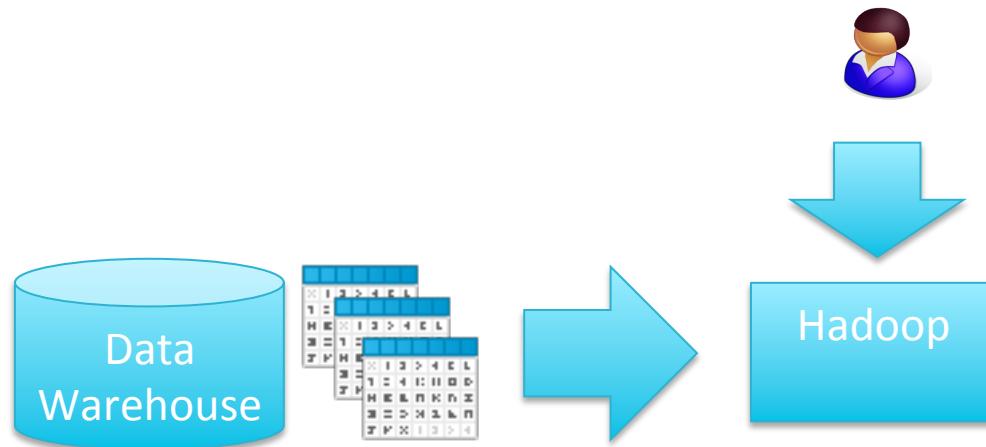
Other Ways Hadoop is Being Leveraged

# Data Archiving Before Hadoop

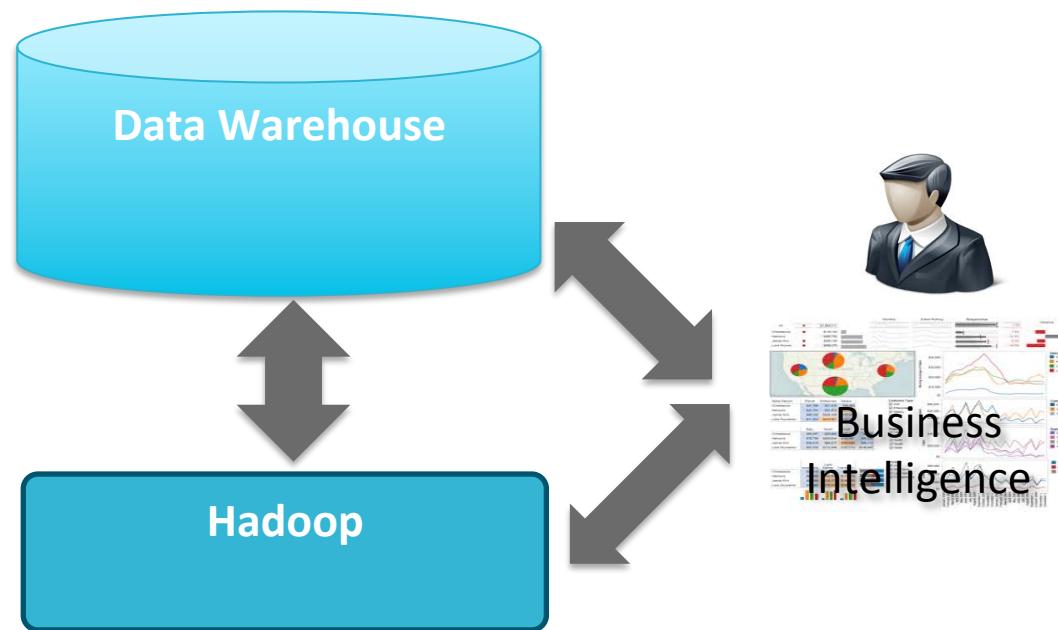
---



# Active Archiving with Hadoop

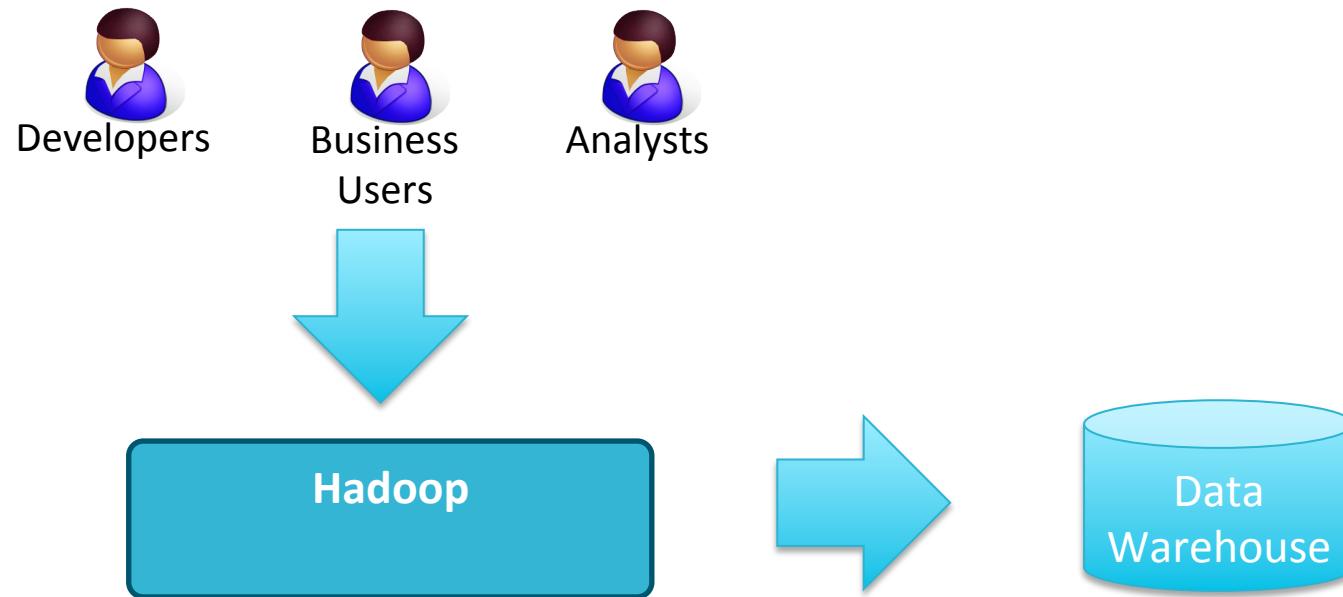


# Offloading Analysis



# Exploratory Analysis

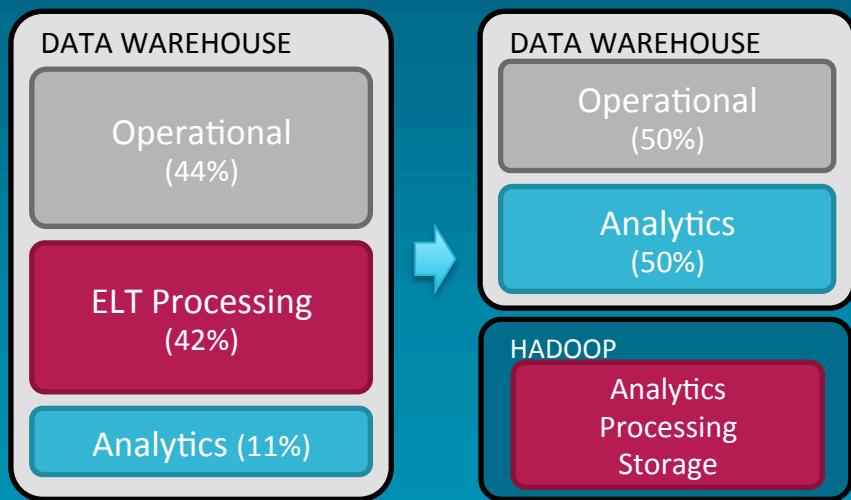
---



# Use Case: A Major Financial Institution

## The Challenge:

- Current EDW at capacity; cannot support growing data depth and width
- Performance issues in business critical apps; little room for innovation.

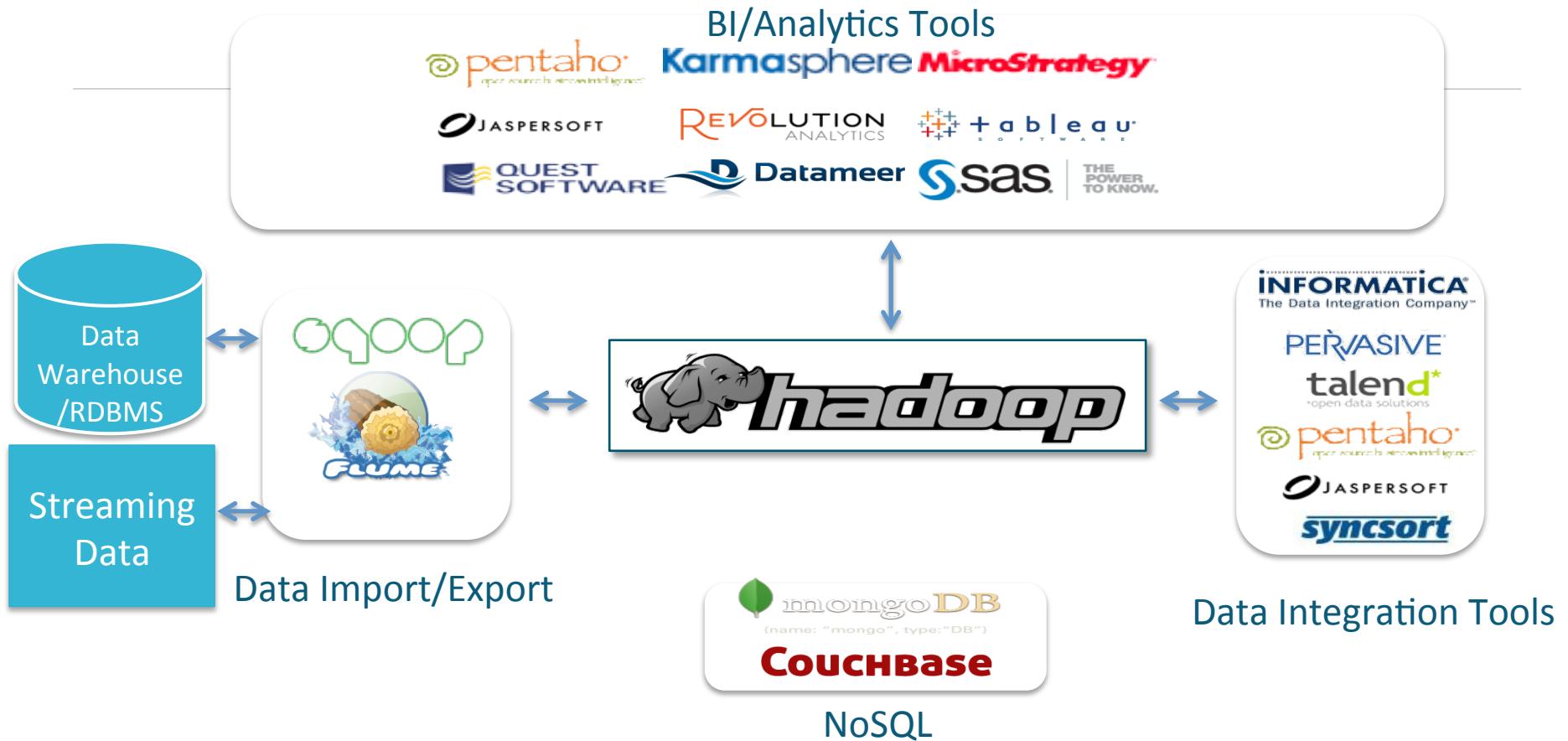


## The Solution:

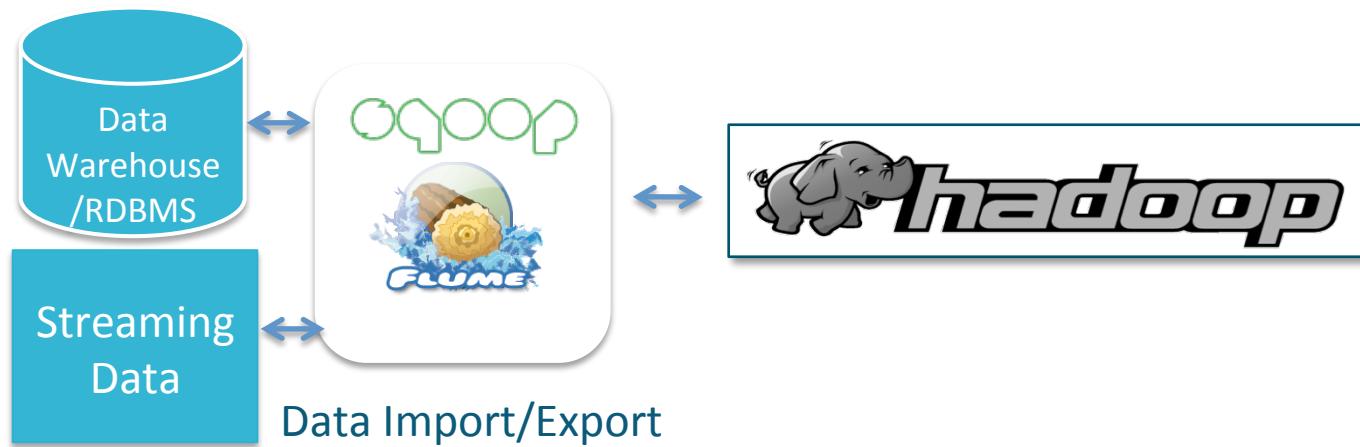
- Hadoop offloads data storage (S), processing (T) & some analytics (Q) from the EDW.
- EDW resources can now be focused on repeatable operational analytics.
- Month data scan in 4 secs vs. 4 hours

# Hadoop Integration

The Big Picture



# Data Import/Export Tools



# Flume in 2 Minutes

---

Or, why you shouldn't be using scripts for data movement.

- Reliable, distributed, and available system for efficient collection, aggregation and movement of streaming data, e.g. logs.
- Open-source, Apache project.

# Sqoop Overview

---

- Apache project designed to ease import and export of data between Hadoop and relational databases.
- Provides functionality to do bulk imports and exports of data with HDFS, Hive and HBase.
- Java based. Leverages MapReduce to transfer data in parallel.

# Transformation/Processing

---

- Standard interface is Java MapReduce
- Higher-level interfaces are commonly used:
  - Apache Hive – provides an SQL like interface to data in Hadoop.
  - Apache Pig – declarative language providing functionality to declare a sequence of transformations.
  - Cloudera Impala – real-time SQL query engine on Hadoop
- Both Hive and Pig convert queries into MapReduce jobs and submit to Hadoop for execution.
- Impala has its own execution engine

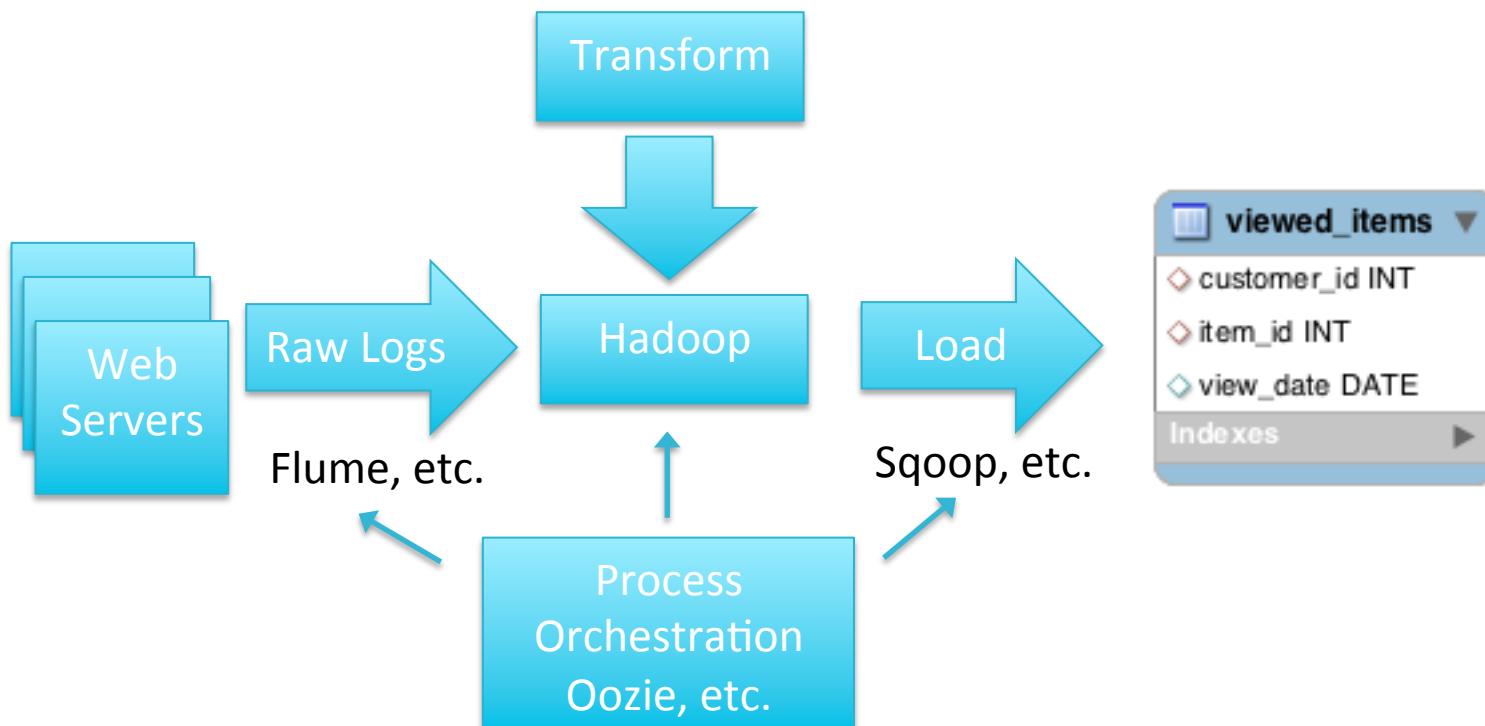
# Orchestration

---

## Schedulers for Hadoop jobs

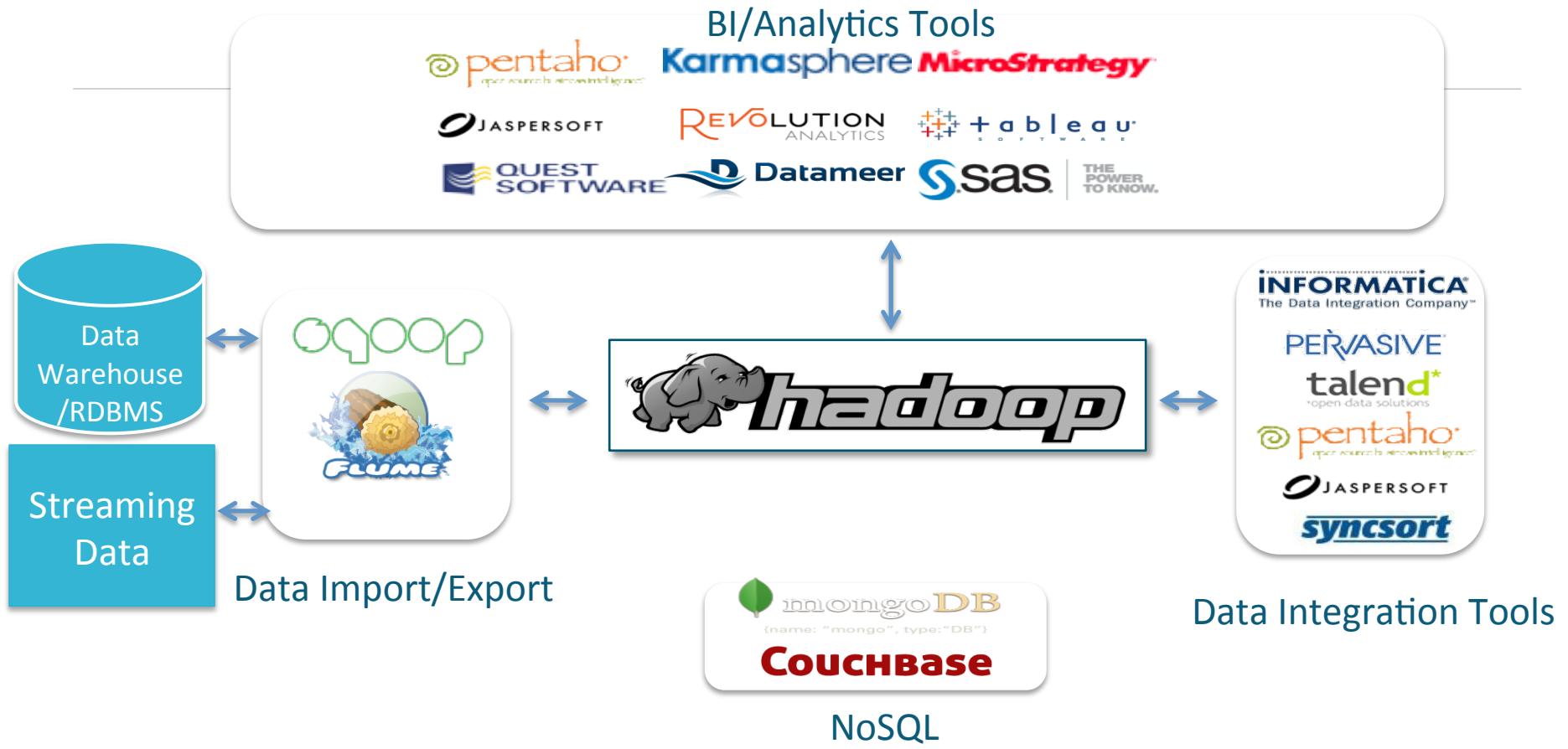
- Oozie
- Azkaban

# Data Flow with OSS Tools



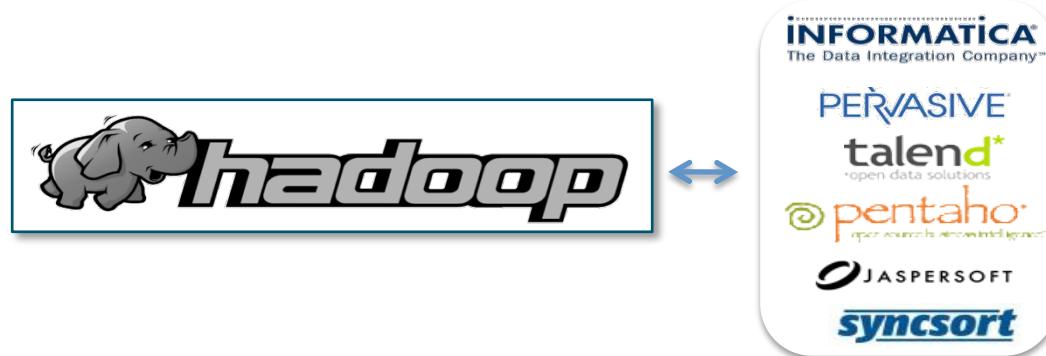
# Hadoop Integration

Data Integration Tools



# Data Integration Tools

---



# Pentaho

---

- Existing BI tools extended to support Hadoop.
- Provides data import/export, transformation, job orchestration, reporting, and analysis functionality.
- Supports integration with HDFS, Hive and HBase.
- Community and Enterprise Editions offered.

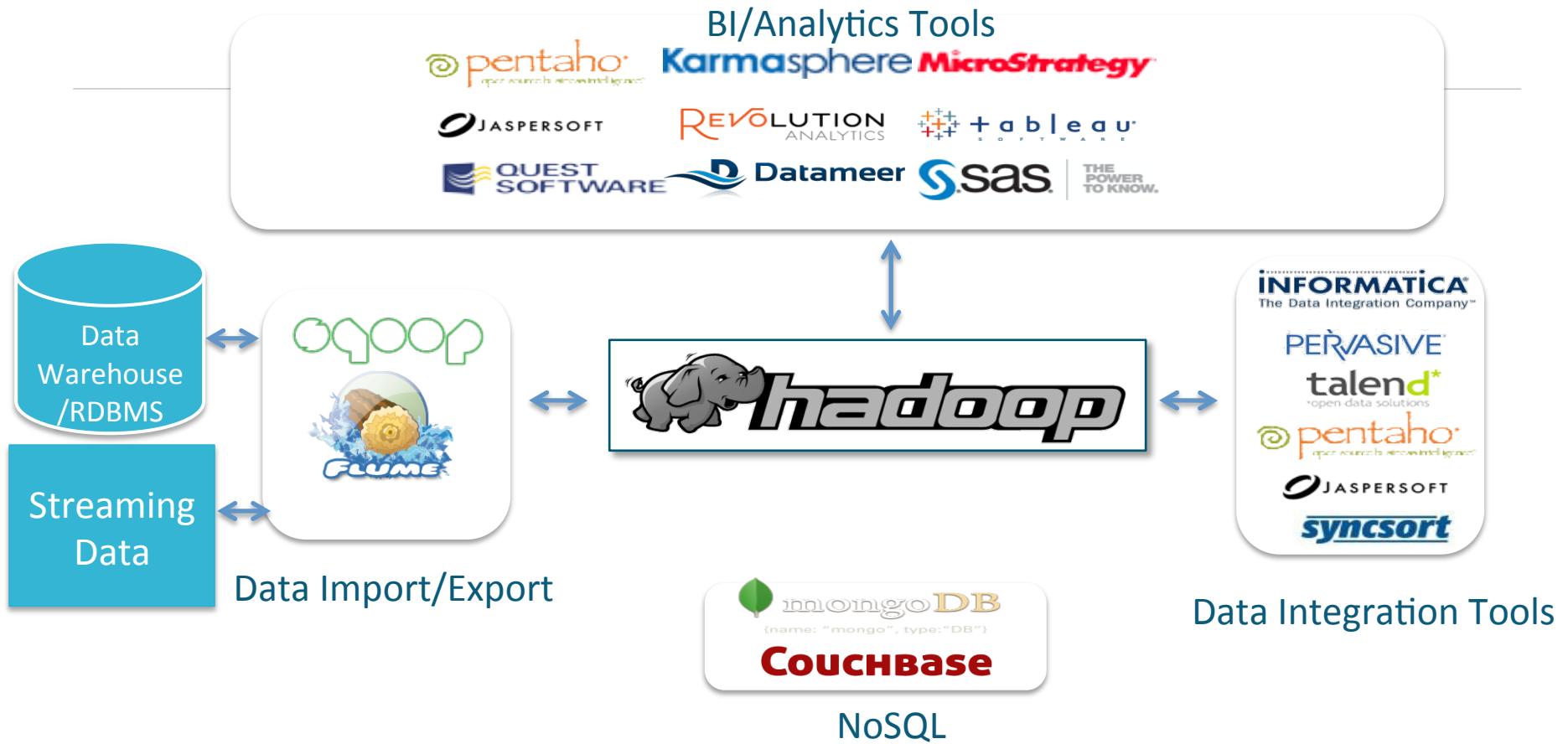
# Informatica

---

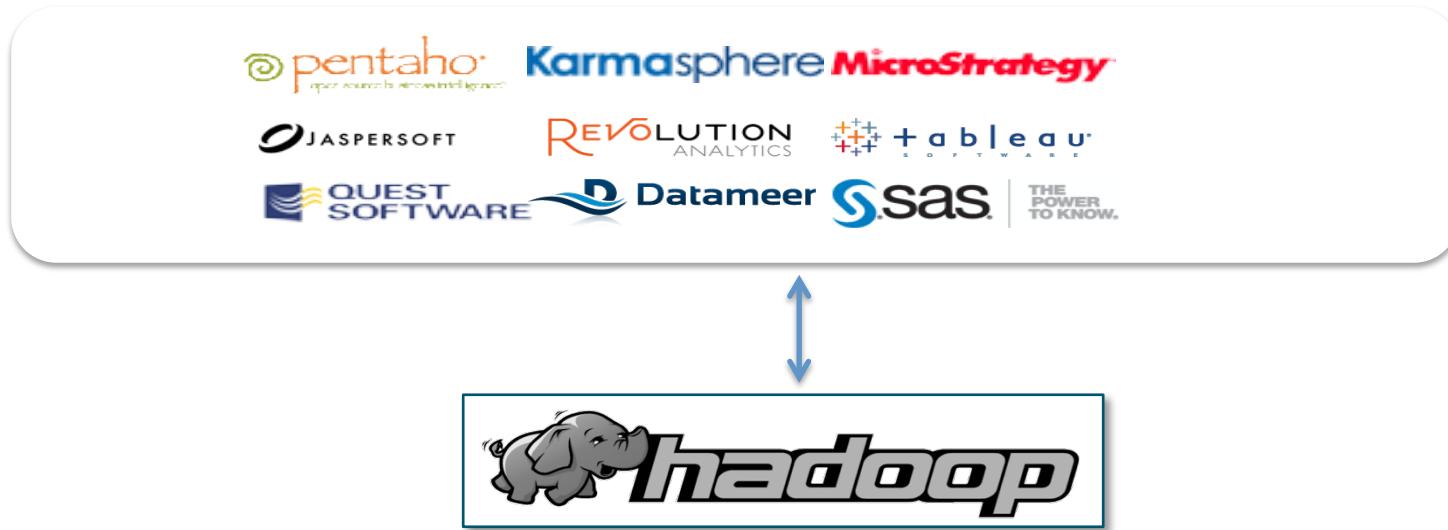
- Informatica
  - Data import/export
  - Metadata services
  - Data lineage
  - Transformation
  - ...

# Hadoop Integration

Business Intelligence/Analytic Tools



# Business Intelligence/Analytics Tools

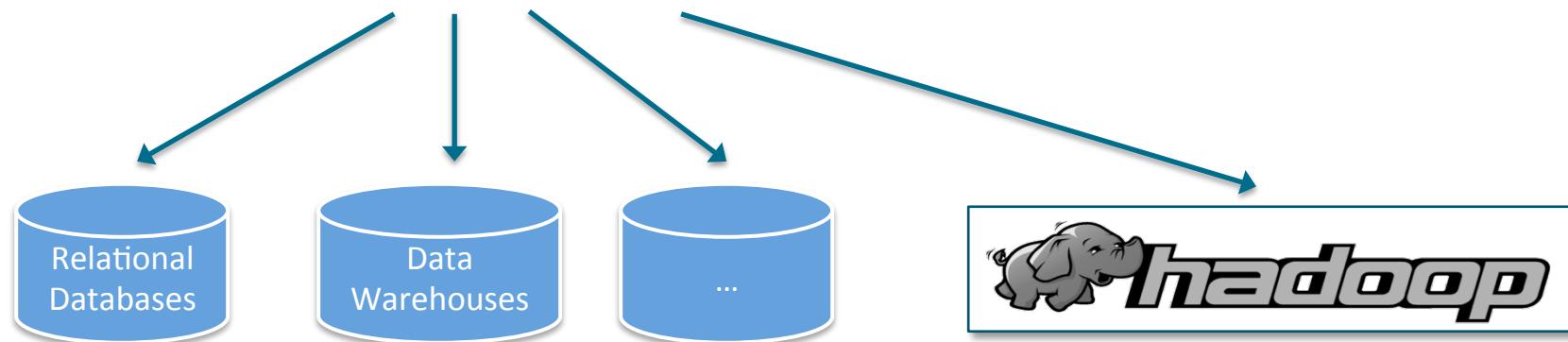


# Business Intelligence/Analytics Tools

MicroStrategy

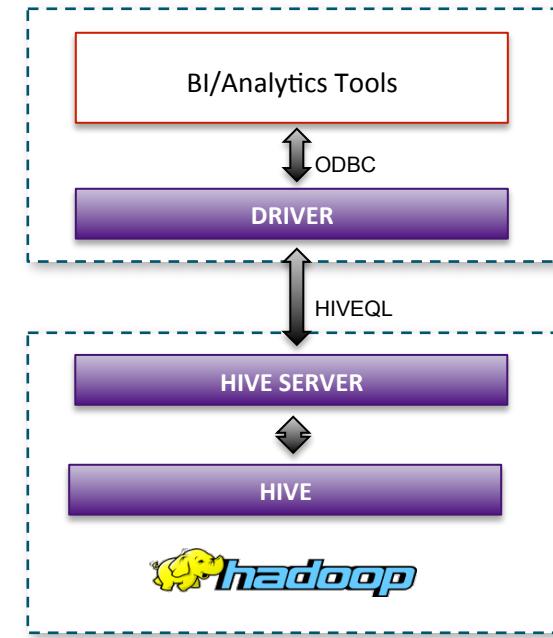
+ tableau  
SOFTWARE

sas | THE POWER TO KNOW.



# ODBC Driver

- Most of these tools use the ODBC standard.
- Since Hive is an SQL-like system it's a good fit for ODBC.
- Several vendors, including Cloudera, make ODBC drivers available for Hadoop.
- JDBC is also used by some products for Hive Integration

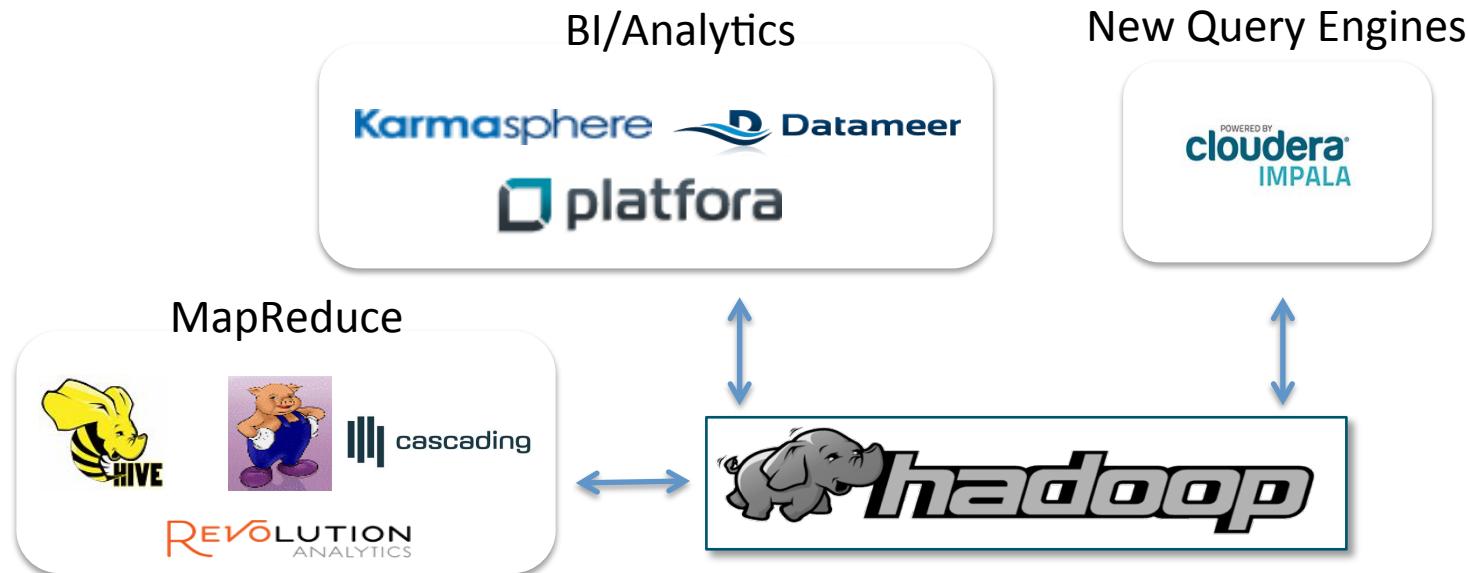


# Hadoop Integration

Next Generation BI/Analytics Tools

# New “Hadoop Native” Tools

You can think of Hadoop as becoming a shared execution environment supporting new data analysis tools...

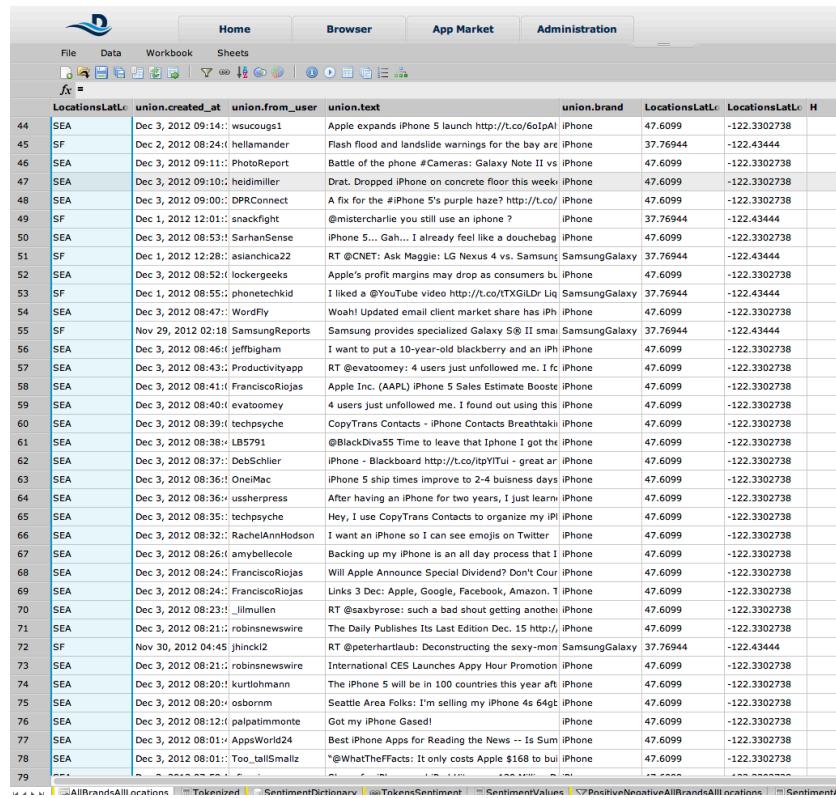


# Hadoop Native Tools – Advantages

---

- New data analysis tools:
  - Designed and optimized for working with Hadoop data and large data sets.
  - Remove reliance on Hive for accessing data – can work with any data in Hadoop.
- New query engines:
  - Provide ability to do low latency queries against Hadoop data.
  - Make it possible to do ad-hoc, exploratory analysis of data in Hadoop.

# Datameer

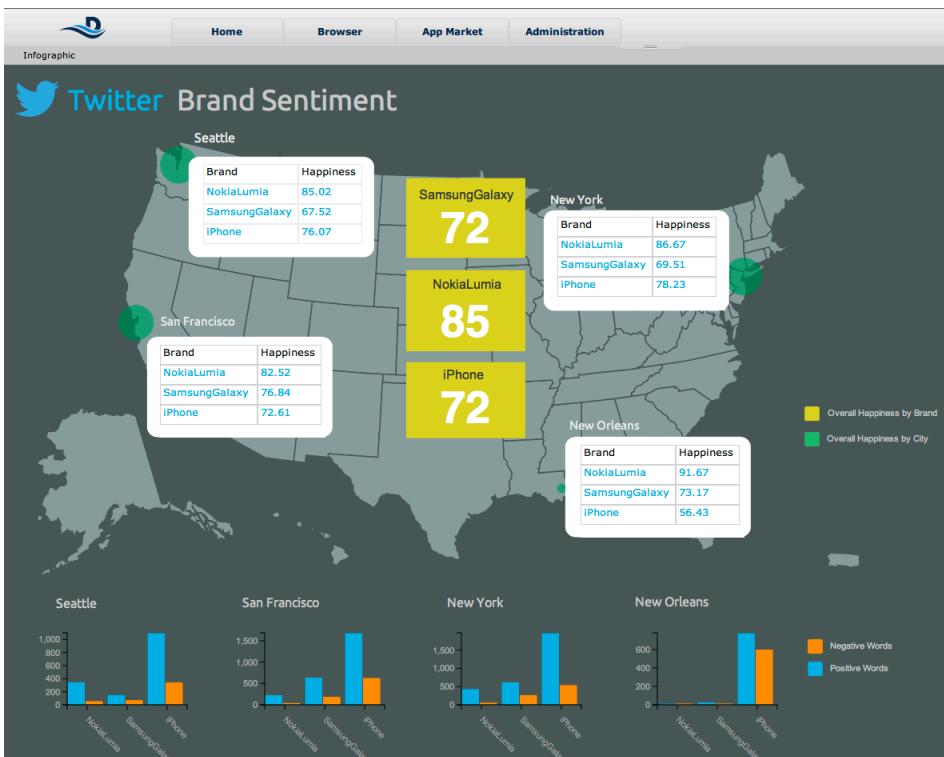


The screenshot shows the Datameer interface with a data browser view. The top navigation bar includes Home, Browser, App Market, and Administration. Below the navigation is a toolbar with various icons for file operations. The main area displays a table of data with the following columns:

|    | LocationsLatLon | union.created_at                    | union.from_user | union.text   | union.brand   | LocationsLatLon | LocationsLatLon | H |
|----|-----------------|-------------------------------------|-----------------|--|---------------|-----------------|-----------------|---|
| 44 | SEA             | Dec 3, 2012 09:14:: wsucougs1       |                 | Apple expands iPhone 5 launch http://t.co/6o1pA1     | iPhone        | 47.6099         | -122.3302738    |   |
| 45 | SF              | Dec 2, 2012 08:24:: hellamander     |                 | Flash flood and landslide warnings for the bay are   | iPhone        | 37.76944        | -122.43444      |   |
| 46 | SEA             | Dec 3, 2012 09:11:: PhotoReport     |                 | Battle of the phone #Cameras: Galaxy Note II vs      | iPhone        | 47.6099         | -122.3302738    |   |
| 47 | SEA             | Dec 3, 2012 09:10:: heidimiller     |                 | Drat. Dropped iPhone on concrete floor this week!    | iPhone        | 47.6099         | -122.3302738    |   |
| 48 | SEA             | Dec 3, 2012 09:00:: DPRConnect      |                 | A fix for the iPhone 5's purple haze? http://t.co/   | iPhone        | 47.6099         | -122.3302738    |   |
| 49 | SF              | Dec 1, 2012 12:01:: snackfight      |                 | @mistercharlie you still use an iphone ?             | iPhone        | 37.76944        | -122.43444      |   |
| 50 | SEA             | Dec 3, 2012 08:53:: SarhanSensis    |                 | iPhone ... Gah... I already feel like a douchebag    | iPhone        | 47.6099         | -122.3302738    |   |
| 51 | SF              | Dec 1, 2012 12:28:: asianchica22    |                 | RT @CNET: Ask Maggie: LG Nexus 4 vs. Samsung         | SamsungGalaxy | 37.76944        | -122.43444      |   |
| 52 | SEA             | Dec 3, 2012 08:52:: lockergEEKs     |                 | Apple's profit margins may drop as consumers b       | iPhone        | 47.6099         | -122.3302738    |   |
| 53 | SF              | Dec 1, 2012 08:55:: phonetechkid    |                 | I liked a YouTube video http://t.co/TxGILDrLqj       | SamsungGalaxy | 37.76944        | -122.43444      |   |
| 54 | SEA             | Dec 3, 2012 08:47:: WordFly         |                 | Woah! Updated email client market share has iPh      | iPhone        | 47.6099         | -122.3302738    |   |
| 55 | SF              | Nov 29, 2012 02:18:: SamsungReports |                 | Samsung provides specialized Galaxy S® II sma        | SamsungGalaxy | 37.76944        | -122.43444      |   |
| 56 | SEA             | Dec 3, 2012 08:46:: jeffbigman      |                 | I want to put a 10-year-old blackberry and an iPh    | iPhone        | 47.6099         | -122.3302738    |   |
| 57 | SEA             | Dec 3, 2012 08:43:: Productivityapp |                 | RT @evatoomey: 4 users just unfollowed me. I fc      | iPhone        | 47.6099         | -122.3302738    |   |
| 58 | SEA             | Dec 3, 2012 08:41:: FranciscoRojas  |                 | App Inc. (AAPL) iPhone 5 Sales Estimate Booste       | iPhone        | 47.6099         | -122.3302738    |   |
| 59 | SEA             | Dec 3, 2012 08:40:: evatoomey       |                 | 4 users just unfollowed me. I found out using thi    | iPhone        | 47.6099         | -122.3302738    |   |
| 60 | SEA             | Dec 3, 2012 08:39:: techpsyche      |                 | CopyTrans Contacts - iPhone Contacts Breathakki      | iPhone        | 47.6099         | -122.3302738    |   |
| 61 | SEA             | Dec 3, 2012 08:38:: LB5791          |                 | @BlackDivas55 Time to leave that Iphone I got th     | iPhone        | 47.6099         | -122.3302738    |   |
| 62 | SEA             | Dec 3, 2012 08:37:: DebSchlier      |                 | iPhone - Blackboard http://t.co/lpYIYtuI - great ar  | iPhone        | 47.6099         | -122.3302738    |   |
| 63 | SEA             | Dec 3, 2012 08:36:: OneiMac         |                 | iPhone 5 ship times improve to 2-4 business days     | iPhone        | 47.6099         | -122.3302738    |   |
| 64 | SEA             | Dec 3, 2012 08:36:: usshepress      |                 | After having an iPhone for 2 years, I just learn     | iPhone        | 47.6099         | -122.3302738    |   |
| 65 | SEA             | Dec 3, 2012 08:35:: techpsyche      |                 | Hey, I use CopyTrans Contacts to organize my iPh     | iPhone        | 47.6099         | -122.3302738    |   |
| 66 | SEA             | Dec 3, 2012 08:32:: RachelAnhHodson |                 | I want an iPhone so I can see emojis on Twitter      | iPhone        | 47.6099         | -122.3302738    |   |
| 67 | SEA             | Dec 3, 2012 08:26:: amybellcole     |                 | Backing up my iPhone is an all day process that i    | iPhone        | 47.6099         | -122.3302738    |   |
| 68 | SEA             | Dec 3, 2012 08:24:: FranciscoRojas  |                 | Will Apple Announce Special Dividend? Don't Cour     | iPhone        | 47.6099         | -122.3302738    |   |
| 69 | SEA             | Dec 3, 2012 08:24:: FranciscoRojas  |                 | Links 3 Dec: Apple, Google, Facebook, Amazon, T      | iPhone        | 47.6099         | -122.3302738    |   |
| 70 | SEA             | Dec 3, 2012 08:23:: _lilmullen      |                 | RT @axazyrose: such a bad about getting another      | iPhone        | 47.6099         | -122.3302738    |   |
| 71 | SEA             | Dec 3, 2012 08:21:: robbinsnewswire |                 | The Daily Publishes Its Last Edition Dec. 15 http:// | iPhone        | 47.6099         | -122.3302738    |   |
| 72 | SF              | Nov 30, 2012 04:45:: jhinchik2      |                 | RT @peterhartlaub: Deconstructing the sexy-mon       | SamsungGalaxy | 37.76944        | -122.43444      |   |
| 73 | SEA             | Dec 3, 2012 08:21:: robbinsnewswire |                 | International CES Launches Appy Hour Promotion       | iPhone        | 47.6099         | -122.3302738    |   |
| 74 | SEA             | Dec 3, 2012 08:20:: kurtlothmann    |                 | The iPhone 5 will be in 100 countries this year aft  | iPhone        | 47.6099         | -122.3302738    |   |
| 75 | SEA             | Dec 3, 2012 08:20:: osbornm         |                 | Seattle Area Folks: I'm selling my iPhone 4s 64GB    | iPhone        | 47.6099         | -122.3302738    |   |
| 76 | SEA             | Dec 3, 2012 08:12:: palpatimmonite  |                 | Got my iPhone Gased!                                 | iPhone        | 47.6099         | -122.3302738    |   |
| 77 | SEA             | Dec 3, 2012 08:01:: AppsWorld24     |                 | Best iPhone Apps for Reading the News -- Is Sum      | iPhone        | 47.6099         | -122.3302738    |   |
| 78 | SEA             | Dec 3, 2012 08:01:: Too_tallSmalz   |                 | @WhatTheFacts: It only costs Apple \$168 to bui      | iPhone        | 47.6099         | -122.3302738    |   |
| 79 | SEA             | Dec 3, 2012 08:01:: Too_tallSmalz   |                 | Phone  | iPhone        | 47.6099         | -122.3302738    |   |

AllBrandsAll locations Tokenized SentimentDictionary TokensSentiment SentimentValues PositiveNegativeAllBrandsAll locations SentimentAll

# Datameer



# Limitations of MapReduce

---

- Queries are slow
- Need a higher level, more real-time abstraction

# New Query Engines – Impala

---

- Fast, interactive queries on data stored in Hadoop (HDFS and HBase).
  - But also designed to support long running queries.
- Uses familiar Hive Query Language and shares metastore.
- Tight integration with Hadoop.
  - Reads common Hadoop file formats.
  - Runs on Hadoop DataNodes.
- High Performance
  - C++, not Java.
  - Runtime code generation.
  - Entirely re-designed execution engine bypasses MapReduce.

# Impala Architecture

Common Hive SQL and interface



Unified metadata and scheduler



Query Planner

Query Coordinator

Query Exec Engine

HDFS DN

HBase

Query Planner

Query Coordinator

Query Exec Engine

HDFS DN

HBase

Query Planner

Query Coordinator

Query Exec Engine

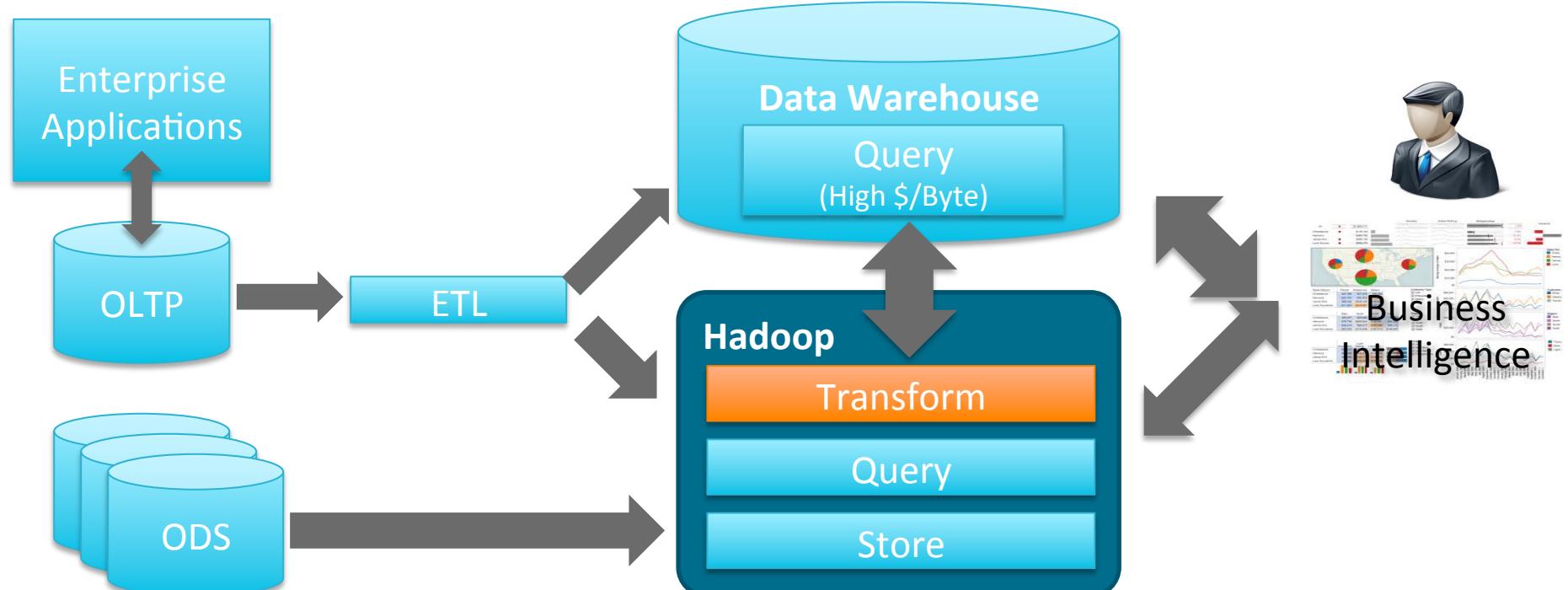
HDFS DN

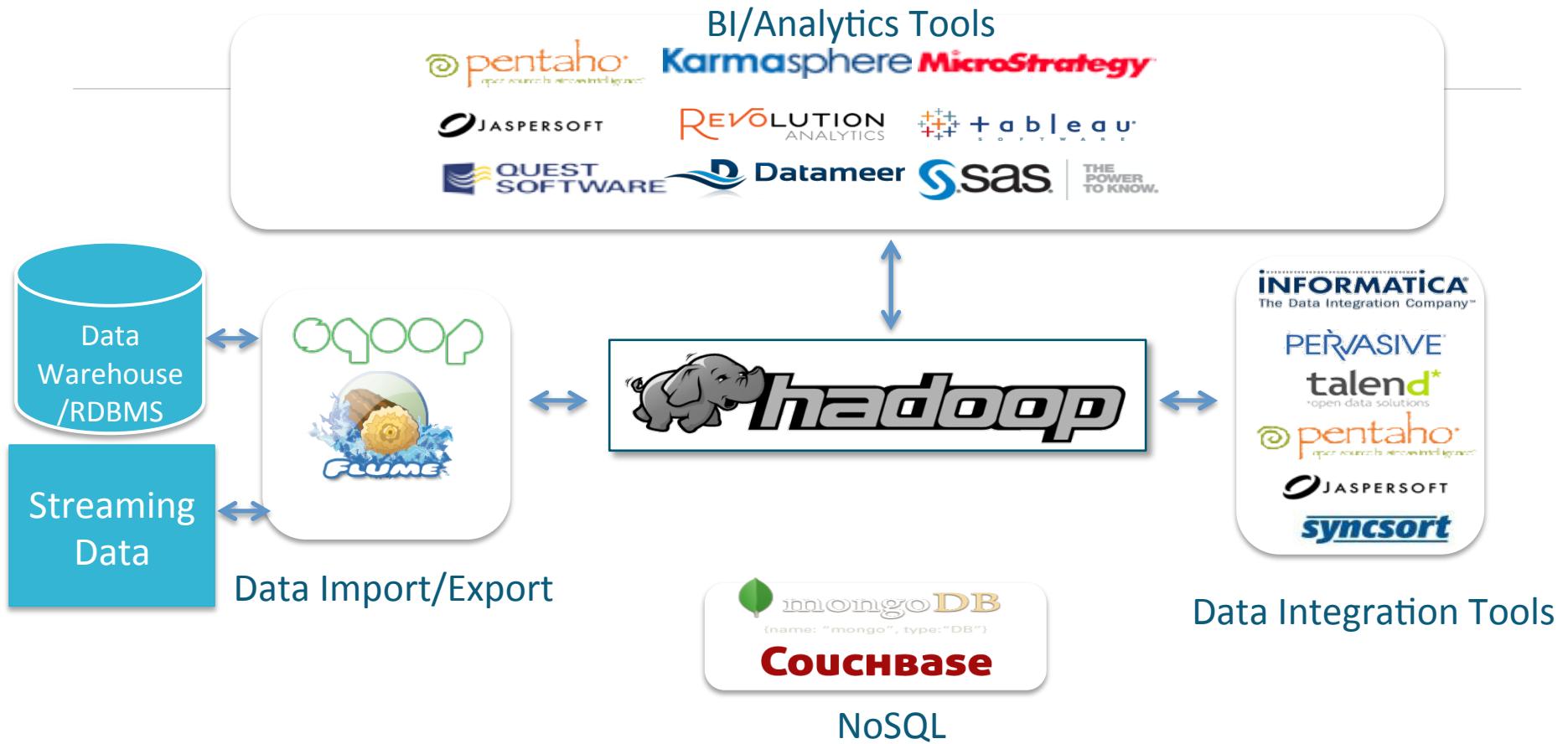
HBase

Fully MPP  
Distributed

# Recap

# Data Warehouse Optimization



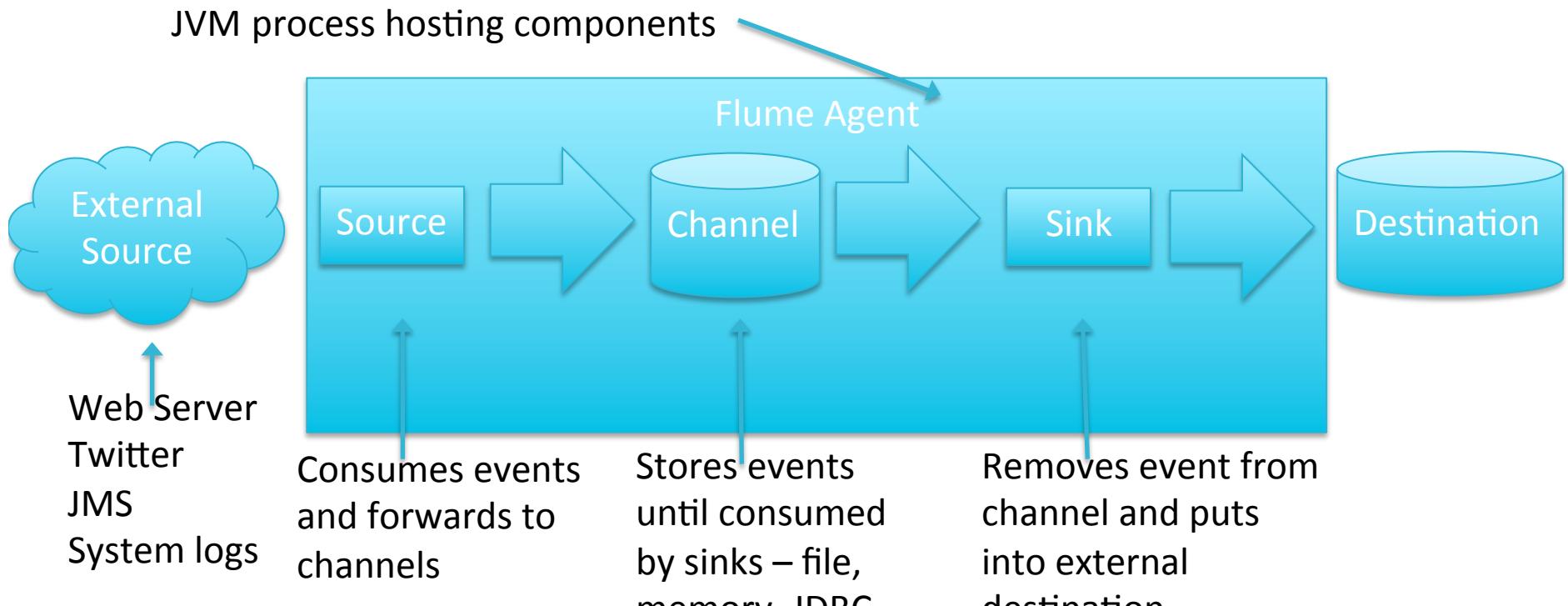


# Questions?

---

- [linkedin.com/in/grover\\_mark](https://linkedin.com/in/grover_mark)
- [mgrover@cloudera.com](mailto:mgrover@cloudera.com)
- Thanks to Jonathan Seidman for sharing his slides

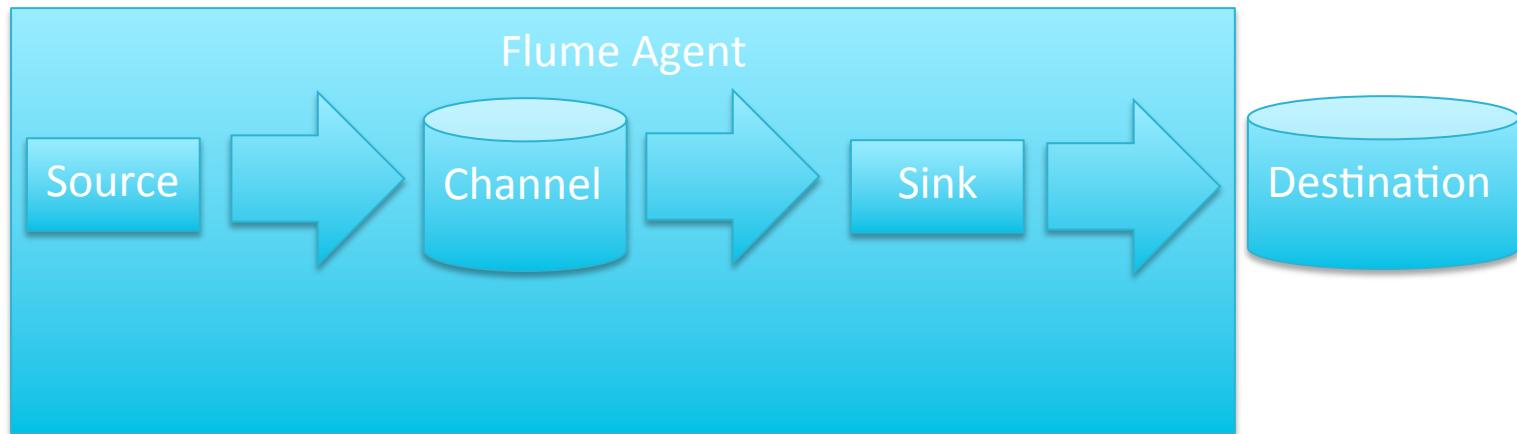
# Flume in 2 Minutes



# Flume in 2 Minutes

---

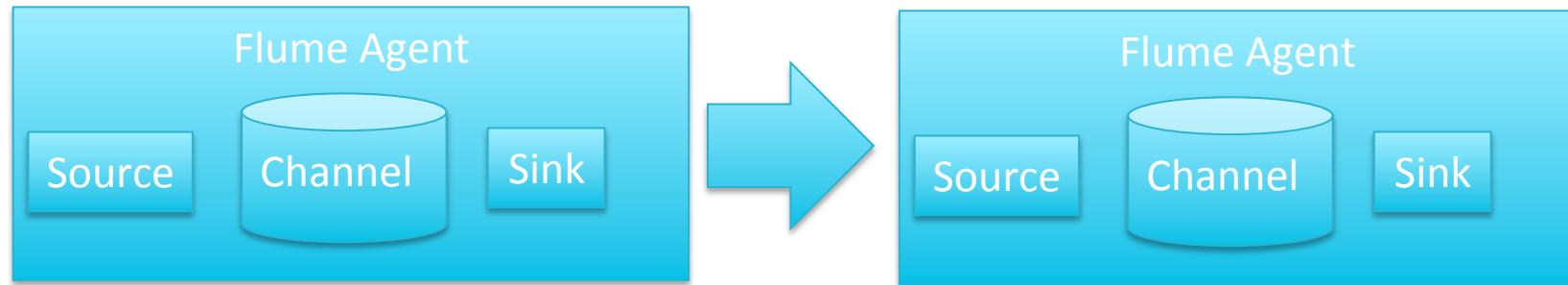
- Reliable – events are stored in channel until delivered to next stage.
- Recoverable – events can be persisted to disk and recovered in the event of failure.



# Flume in 2 Minutes

---

- Supports multi-hop flows for more complex processing.
- Also fan-out, fan-in.



# Flume in 2 Minutes

---

- Declarative
  - No coding required.
  - Configuration specifies how components are wired together.

```
# example.conf: A single-node Flume configuration

# Name the components on this agent
a1.sources = r1
a1.sinks = k1
a1.channels = c1

# Describe/configure the source
a1.sources.r1.type = netcat
a1.sources.r1.bind = localhost
a1.sources.r1.port = 44444

# Describe the sink
a1.sinks.k1.type = logger

# Use a channel which buffers events in memory
a1.channels.c1.type = memory
a1.channels.c1.capacity = 1000
a1.channels.c1.transactionCapacity = 100

# Bind the source and sink to the channel
a1.sources.r1.channels = c1
a1.sinks.k1.channel = c1
```

# Flume in 2 Minutes

---

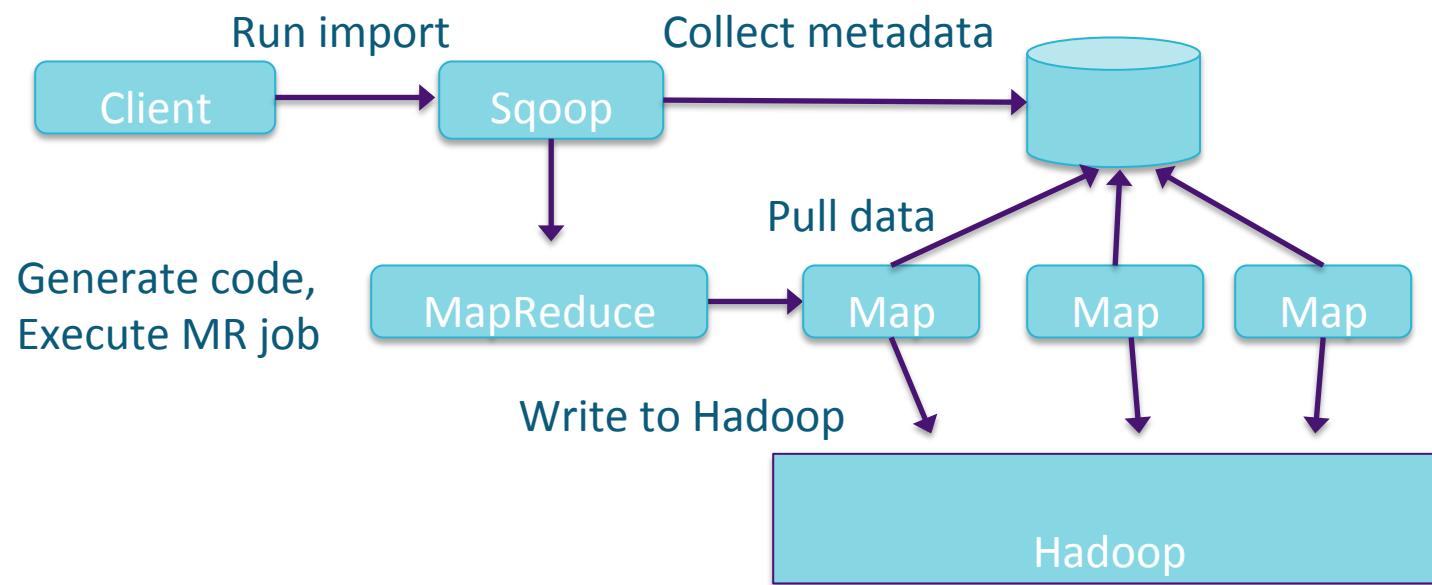
- Similar systems:
  - Scribe
  - Chukwa

# Sqoop Overview

---

- Uses a “connector” abstraction.
- Two types of connectors
  - Standard connectors are JDBC based.
  - Direct connectors use native database interfaces to improve performance.
- Direct connectors are available for many open-source and commercial databases – MySQL, PostgreSQL, Oracle, SQL Server, Teradata, etc.

# Sqoop Import Flow



# Sqoop Limitations

---

Sqoop has some limitations, including:

- Poor support for security.  
    \$ sqoop import --username scott --password tiger...
- Sqoop can read command line options from an option file, but this still has holes.
- Error prone syntax.
- Tight coupling to JDBC model – not a good fit for non-RDBMS systems.

## Fortunately...

---

Sqoop 2 (incubating) will address many of these limitations:

- Adds a web-based GUI.
- Centralized configuration.
- More flexible model.
- Improved security model.

# Hive Integration

---

## HiveServer1

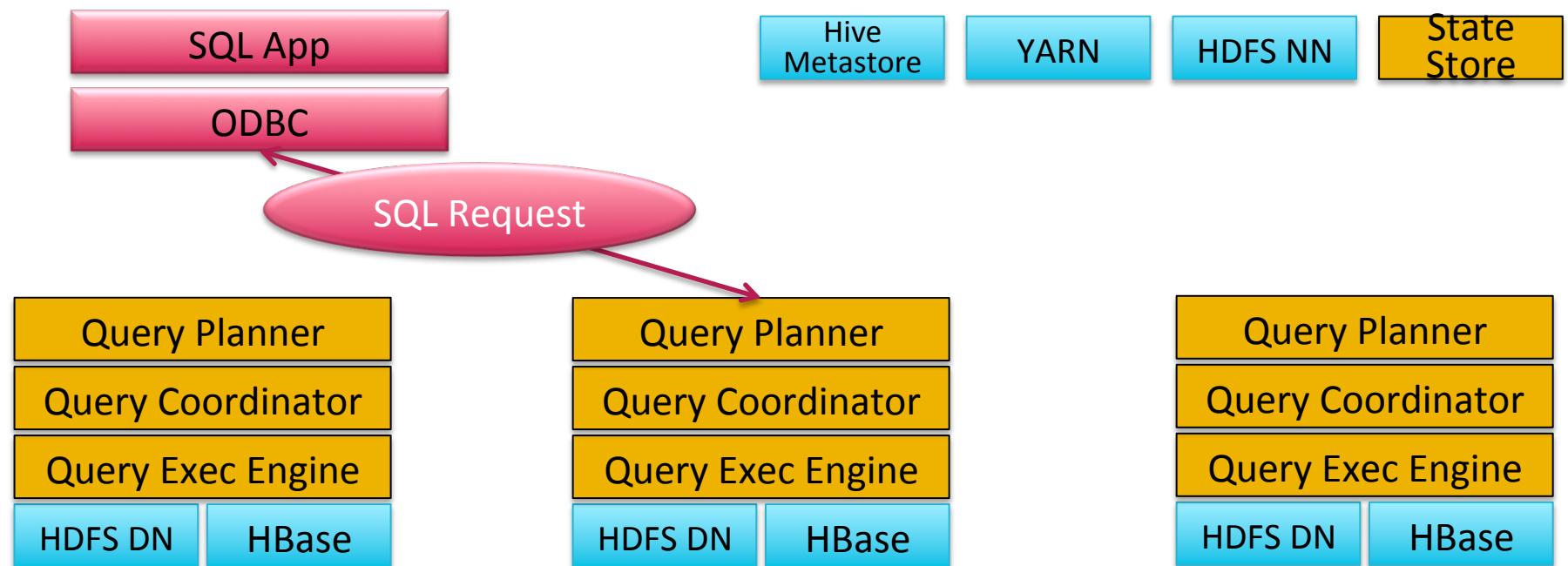
- No support for concurrent queries. Requires running multiple HiveServers for multiple users
- No support for security.
- The Thrift API in the Hive Server doesn't support common JDBC/ODBC calls.

## HiveServer2

- Adds support for concurrent queries. Can support multiple users.
- Adds security support with Kerberos.
- Better support for JDBC and ODBC.

# Cloudera Impala Details

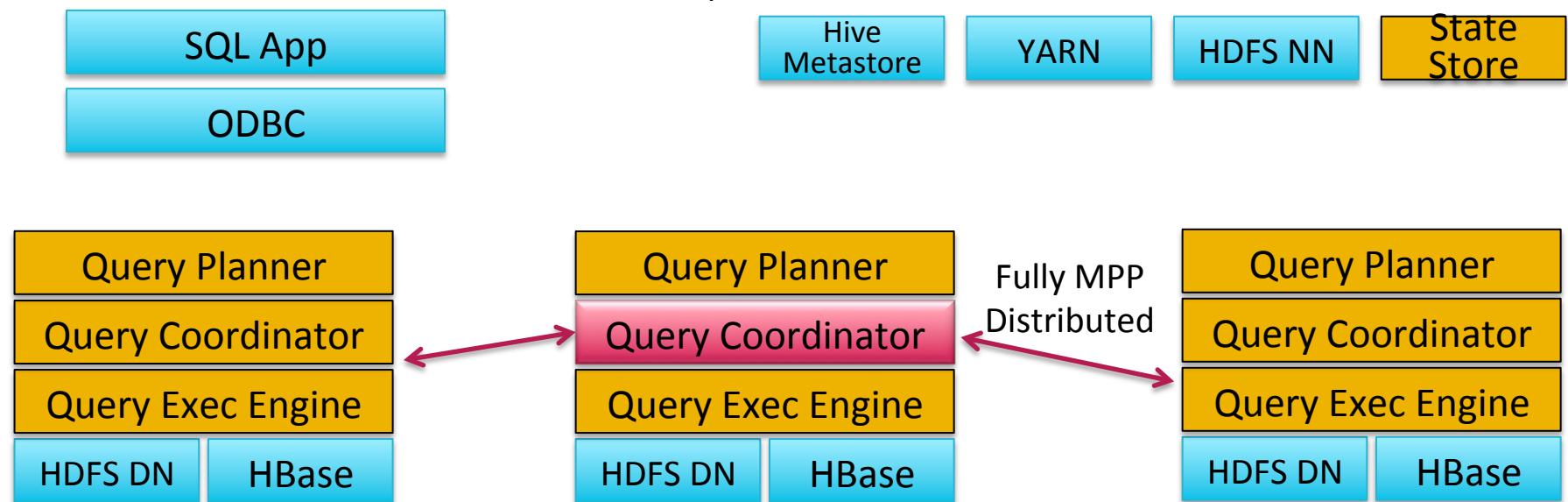
Client submits query through ODBC



# Cloudera Impala Details

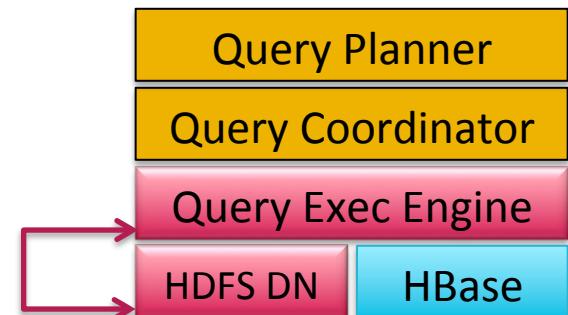
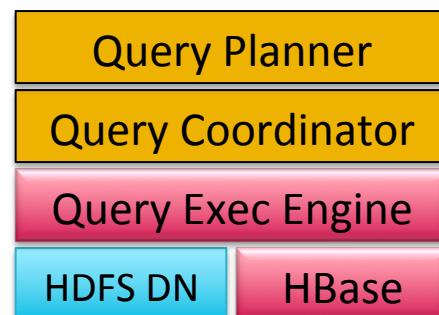
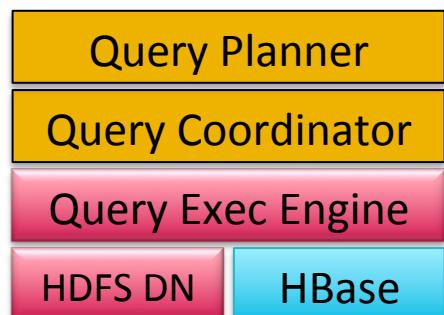
Planner turns request into collection of plan fragments.

Coordinator initiates execution on remote impalad's



# Cloudera Impala Details

Impalads participating in query access local data in HDFS or HBase



Local Direct Reads

# Cloudera Impala Details

Intermediate results are streamed between impalad's  
Final results are streamed back to client

