Big Data Retrospective

STL Big Data IDEA

January 2019

Agenda

- Introduction
- Continue
- Stop
- Start
- Questions

Our Speakers

Adam Doyle

David Youngberg

Scott Shaw

Introduction

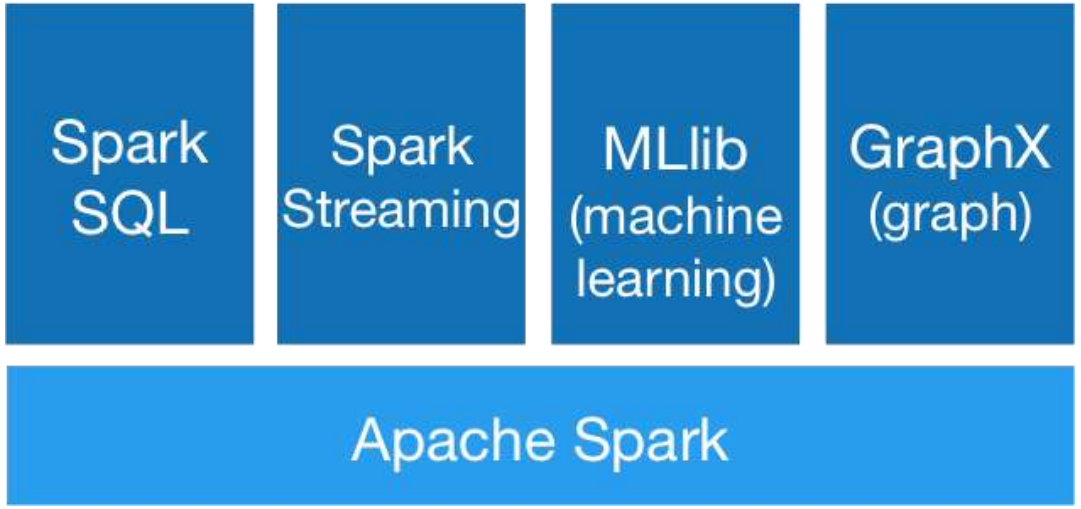
- Sprint Retrospective
  - Opportunity for the group to inspect itself and create a plan for improvements to be enacted in the next Sprint.

  - Things that are going well – Continue
  - Things that could be improved – Stop
  - Things to work on in this coming sprint – Start

- Applying to the Big Data world

CONTINUE

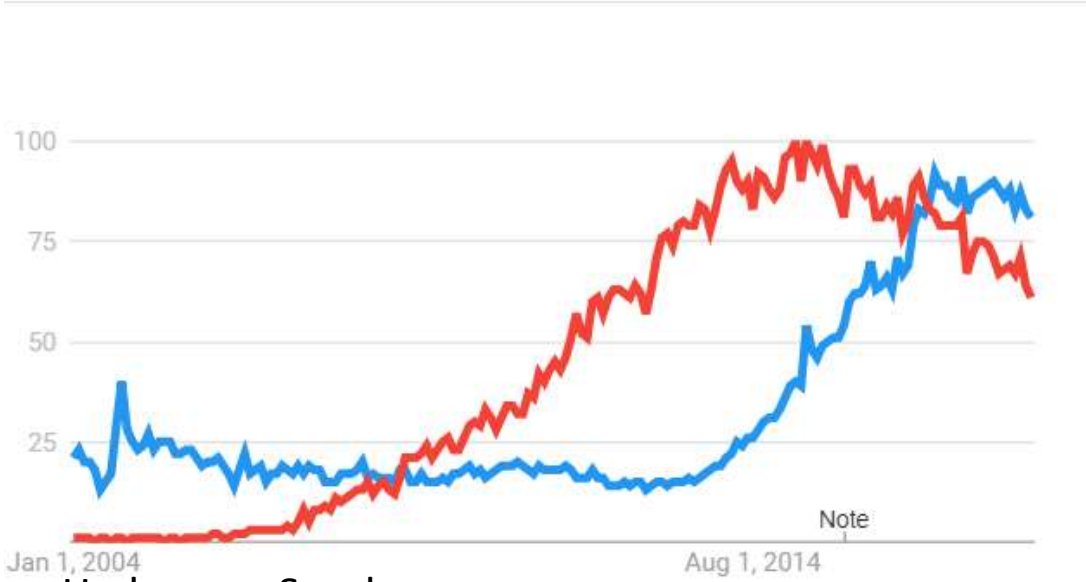
Spark




```
df = spark.read.json("logs.json")
df.where("age > 21")  .select("name.first").show()
```

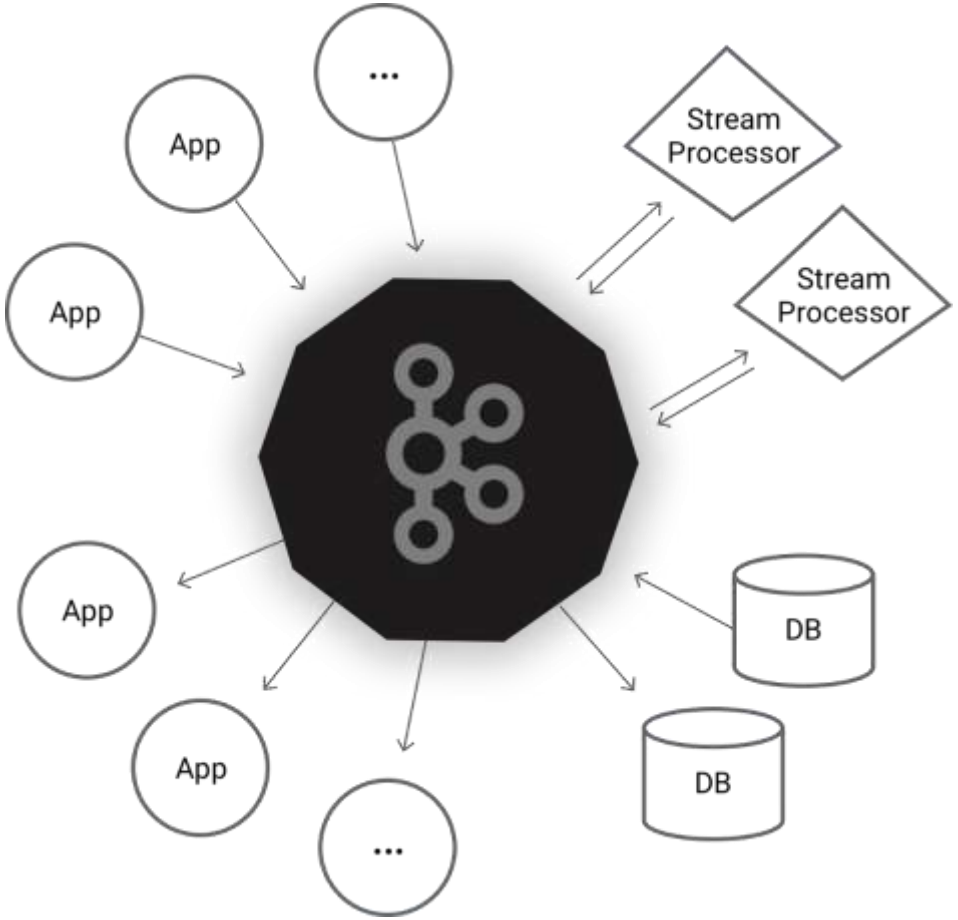
Spark

- Major player in the Big Data ecosystem
- Shift from storage to computational power
- Improved cloud-based infrastructures
- Improved security and governance models



Hadoop vs Spark

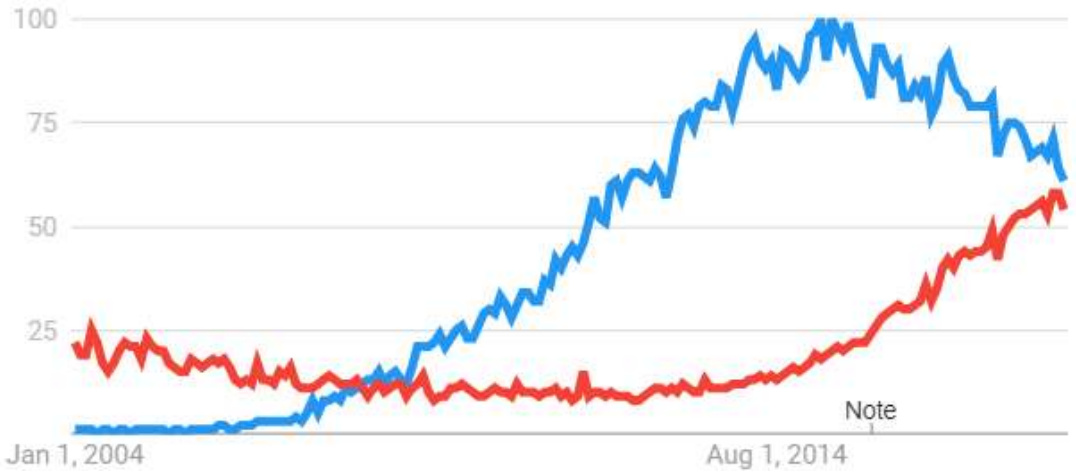
Kafka



Kafka

Confluent is developing Kafka beyond just a mechanism for buffering streaming data

- Kafka Connect
- Kafka Streams
- KSQL
- Schema Registry



Hadoop vs Kafka

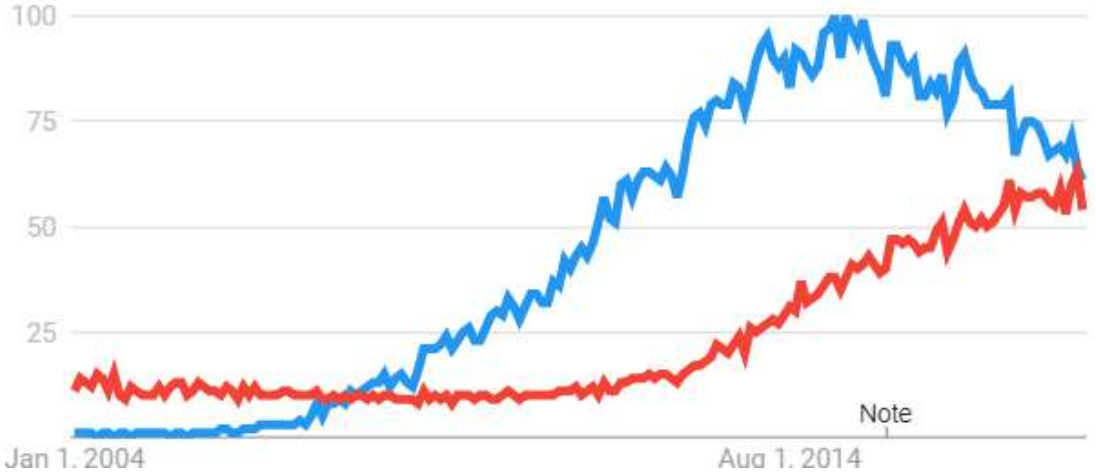
Hive


```
CREATE [TEMPORARY] [EXTERNAL] TABLE [IF NOT EXISTS] [db_name.]table_name
col_comment], ... [constraint_specification])]
  [COMMENT table_comment]
  [PARTITIONED BY (col_name data_type [COMMENT col_comment], ...)]
  [CLUSTERED BY (col_name, col_name, ...) [SORTED BY (col_name [ASC|DESC], ...)] INTO num_buckets
BUCKETS]
  [SKEWED BY (col_name, col_name, ...)
ON ((col_value, col_value, ...), (col_value, col_value, ...), ...)
     [STORED AS DIRECTORIES]
  [
   [ROW FORMAT row_format]
   [STORED AS file_format]
     | STORED BY 'storage.handler.class.name' [WITH SERDEPROPERTIES (...)]
  ]
  [LOCATION hdfs_path]
  [TBLPROPERTIES (property_name=property_value, ...)]
[AS select_statement];   -- (Note: Available in Hive 0.5.0 and later; not supported for external tables)
```

```
SELECT [ALL | DISTINCT] select_expr, select_expr, ...
  FROM table_reference
  [WHERE where_condition]
  [GROUP BY col_list]
  [ORDER BY col_list]
  [CLUSTER BY col_list
    | [DISTRIBUTE BY col_list] [SORT BY col_list]
  ]
 [LIMIT [offset,] rows]
```

Hive

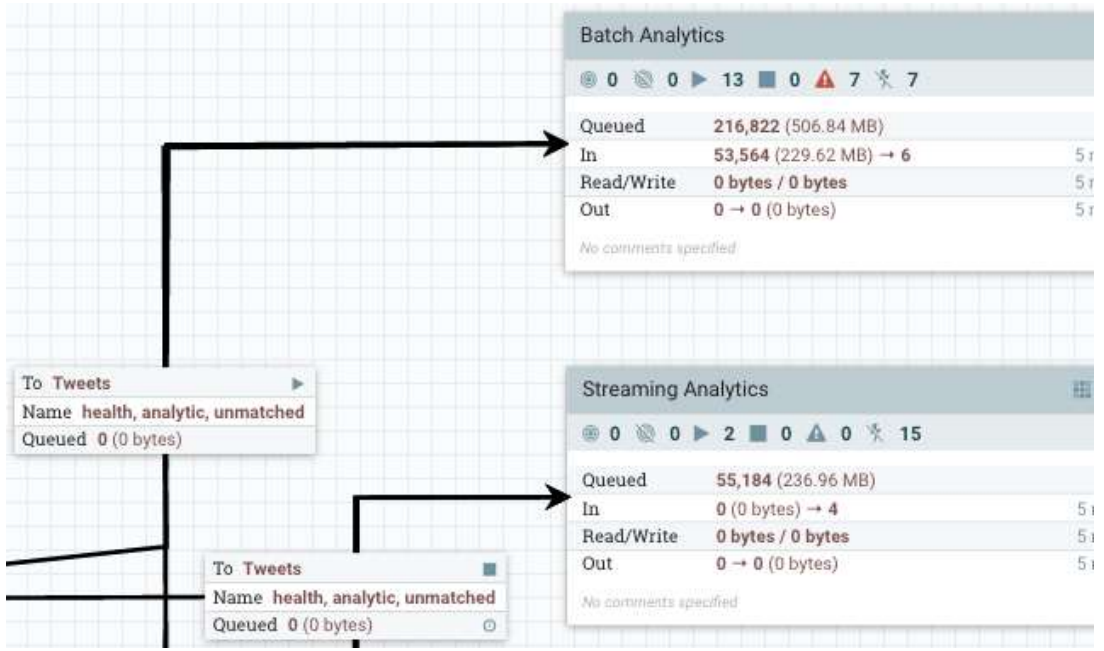
- Other than HDFS, Hive is still the most widely used project in the Hadoop ecosystem.
- Supports schema-on-read as well as structured use cases
- Enhancements to Hive including Tez and LLAP have increased its query speed



Hadoop vs Hive

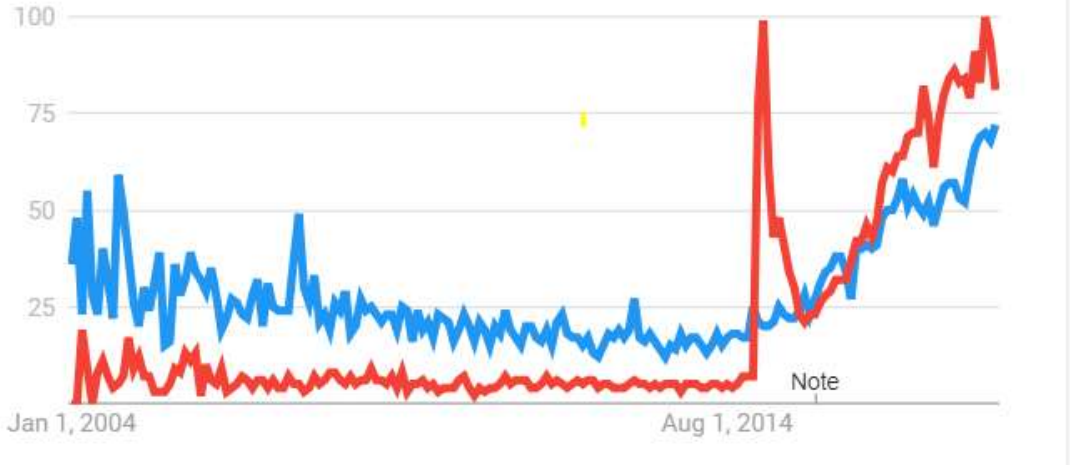
NiFi

- Web-based user interface
  - Seamless experience between design, control, feedback, and monitoring
- Highly configurable
  - Loss tolerant vs guaranteed delivery
  - Low latency vs high throughput
  - Dynamic prioritization
  - Flow can be modified at runtime
  - Back pressure
- Data Provenance
  - Track dataflow from beginning to end
- Designed for extension
  - Build your own processors and more
  - Enables rapid development and effective testing
- Secure
  - SSL, SSH, HTTPS, encrypted content, etc...
  - Multi-tenant authorization and internal authorization/policy management



NiFi

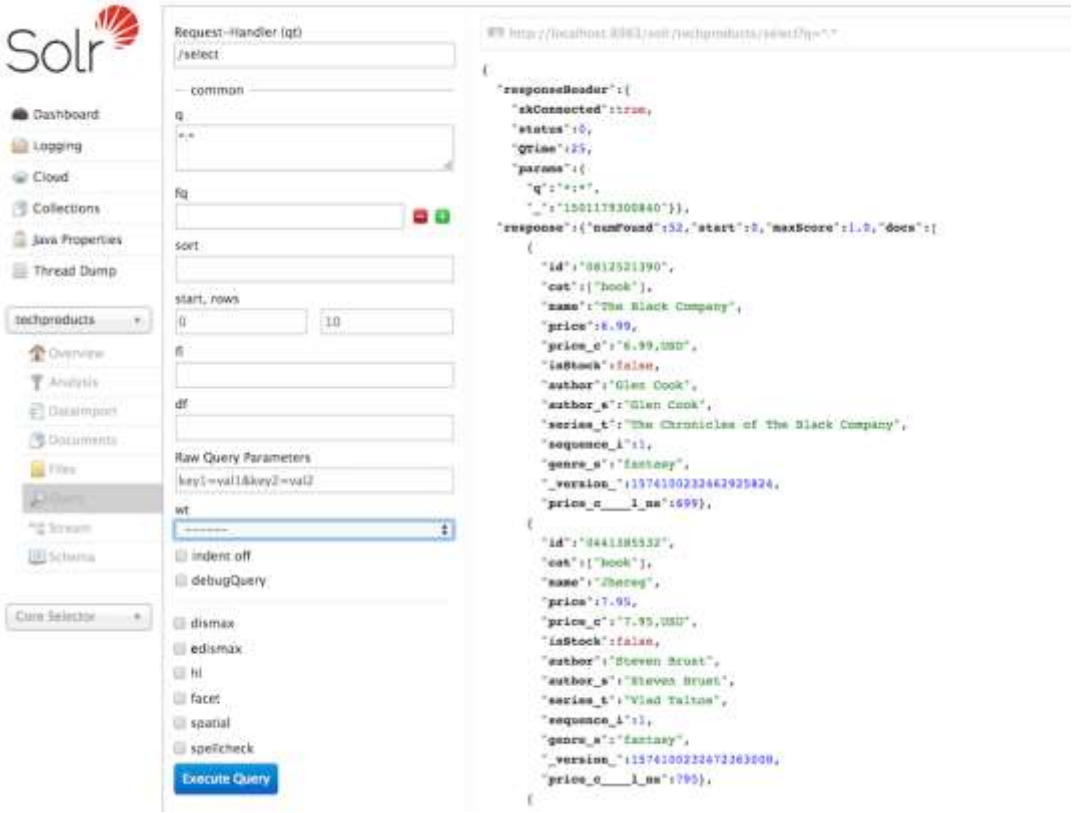
- Tight integrations with data governance platforms
- Data lineage
- Adopted by Hortonworks and Teradata



Flink vs. NiFi

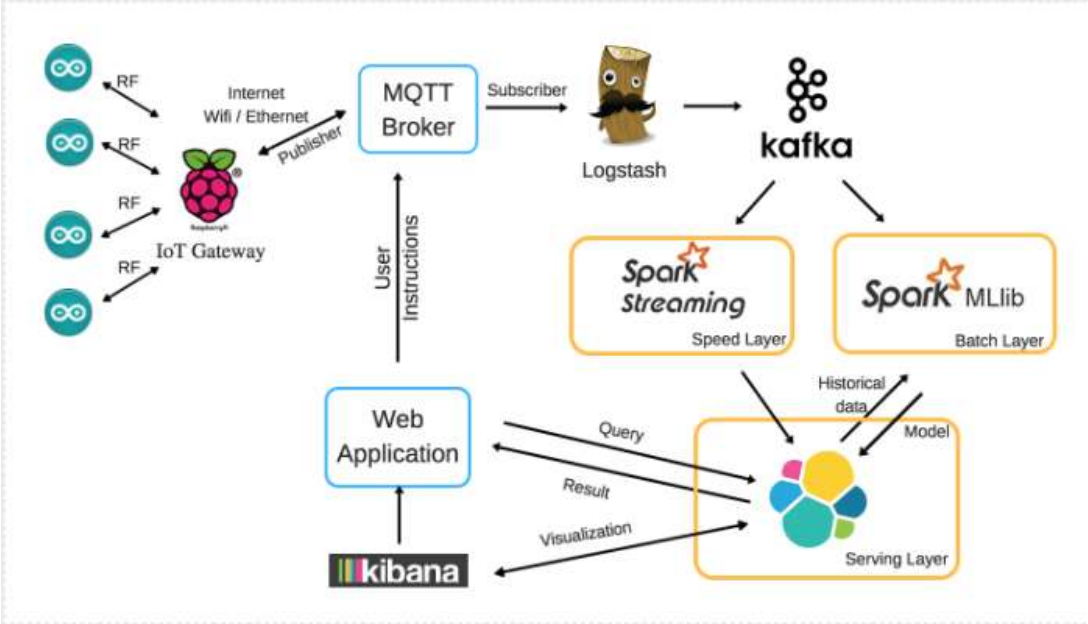
SOLR

- Full-text search
- Optimized for high-traffic
- Near real-time indexing
- Scalable, fault-tolerant



SOLR

- Simplified query language
- Responsive run-time
- Tight integrations with the Cloudera stack
- ELK stack – Elasticsearch, Logstash, Kibana



STOP

Pig

- Release cycles and customer interest slowing down
  - Cloudera CDH 5.X has been on same Pig release (0.12) since 2014
  - Pig team starting to average < 1 release per year
- May not pay to invest time in learning Pig Latin (whole new language) when SQL tools and tools leveraging Java and python exist to do the same work

Oozie

- Unpleasant to use
  - Workflows expressed in XML (not pleasant to hand edit)
  - Reliability issues
- Not the only workflow game on Hadoop
  - Too many to list, Open Source and Commercial products
  - Or even roll your own with Apache Airflow (python DAG library)

Sqoop

- Development slowdown
  - Sqoop 1.X minor releases coming out slowly
  - Sqoop2 still not “prod ready”
- Sqoop gets data out of an RDBMS with limited ability to reformat with command line import options
- Many tools exist that can connect to RDBMS’s and incorporate that connection into a workflow (for example, Nifi, StreamSets, and many others)

Storm

- Does what Spark does, so why not use Spark?
- Commercial Support not available from all Hadoop vendors, unlike Spark.

Flume

- Project Health and Adoption
- Reliability issues
- Many other ecosystem tools out there are ready to receive your streaming data, like Spark Streaming, etc , etc

START

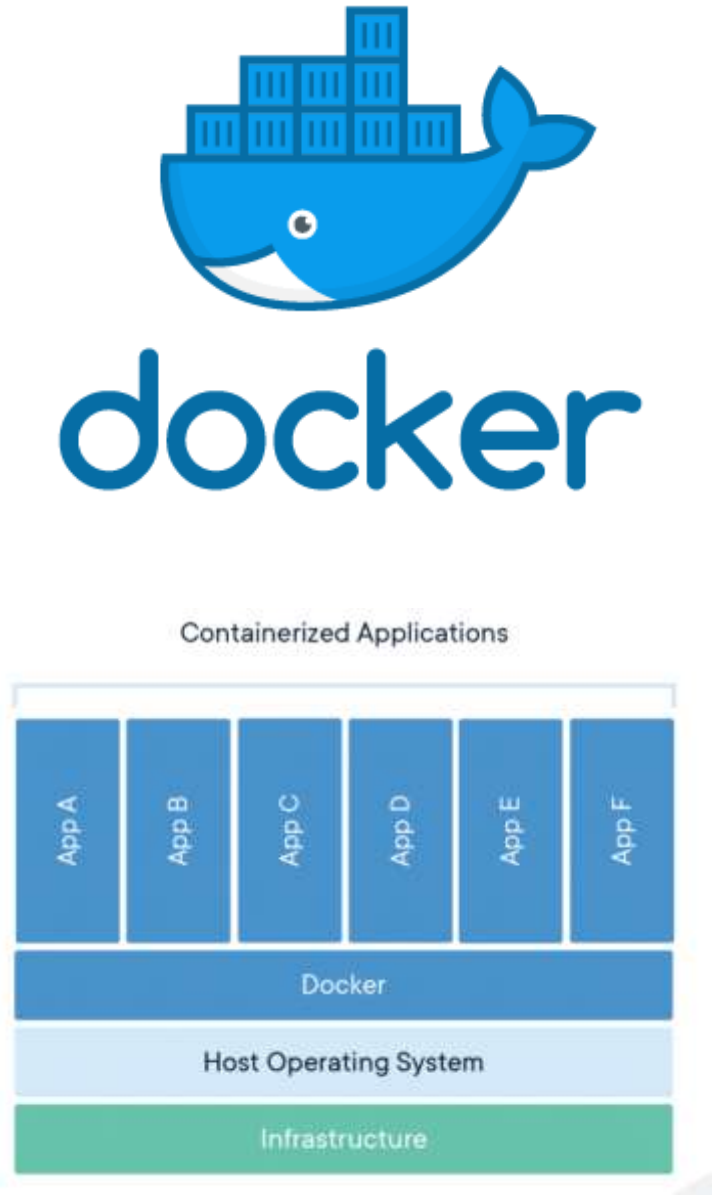
Druid

Druid provides fast analytical queries, at high concurrency, on both real-time and historical data. Druid is often used to power interactive UIs.

Druid is a new type of database that combines ideas from **OLAP/analytic databases**, **timeseries databases**, and **search systems** to enable new use cases in real-time architectures.

Graph Databases

JanusGraph is a scalable graph database optimized for storing and querying graphs containing hundreds of billions of vertices and edges distributed across a multi-machine cluster. JanusGraph is a transactional database that can support thousands of concurrent users executing complex graph traversals in real time.



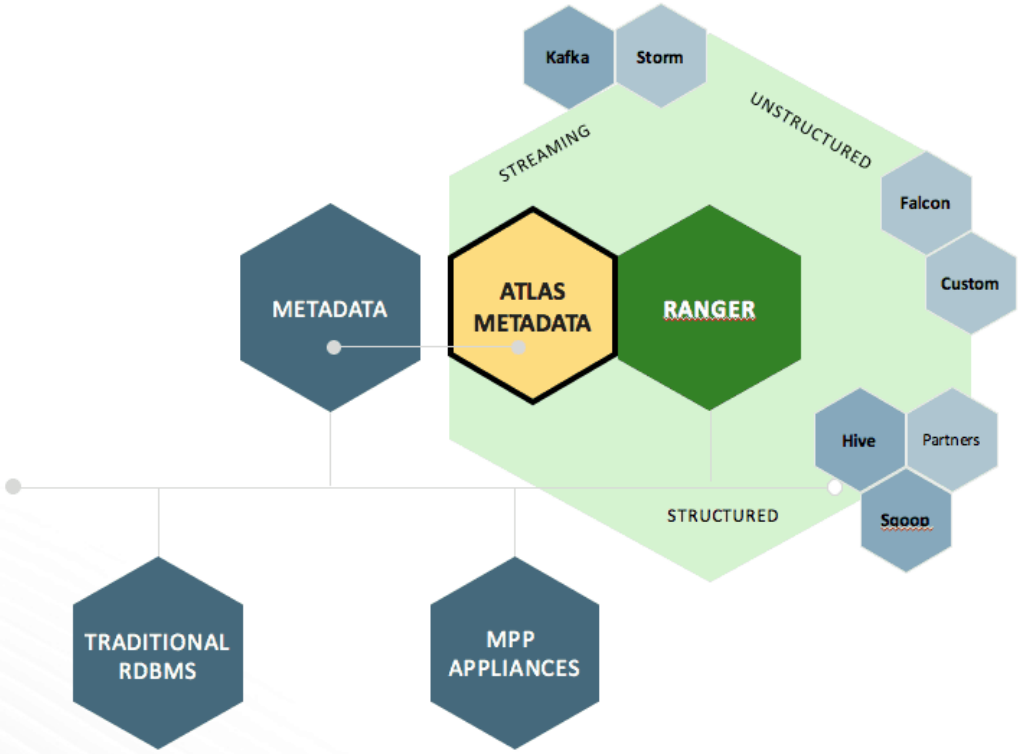
Docker

A container is a standard unit of software that packages up code and all its dependencies so the application runs quickly and reliably from one computing environment to another. A Docker container image is a lightweight, standalone, executable package of software that includes everything needed to run an application: code, runtime, system tools, system libraries and settings.

Atlas

Apache Atlas

Atlas is a scalable and extensible set of core foundational governance services – enabling enterprises to effectively and efficiently meet their compliance requirements within Hadoop and allows integration with the whole enterprise data ecosystem.



Ozone

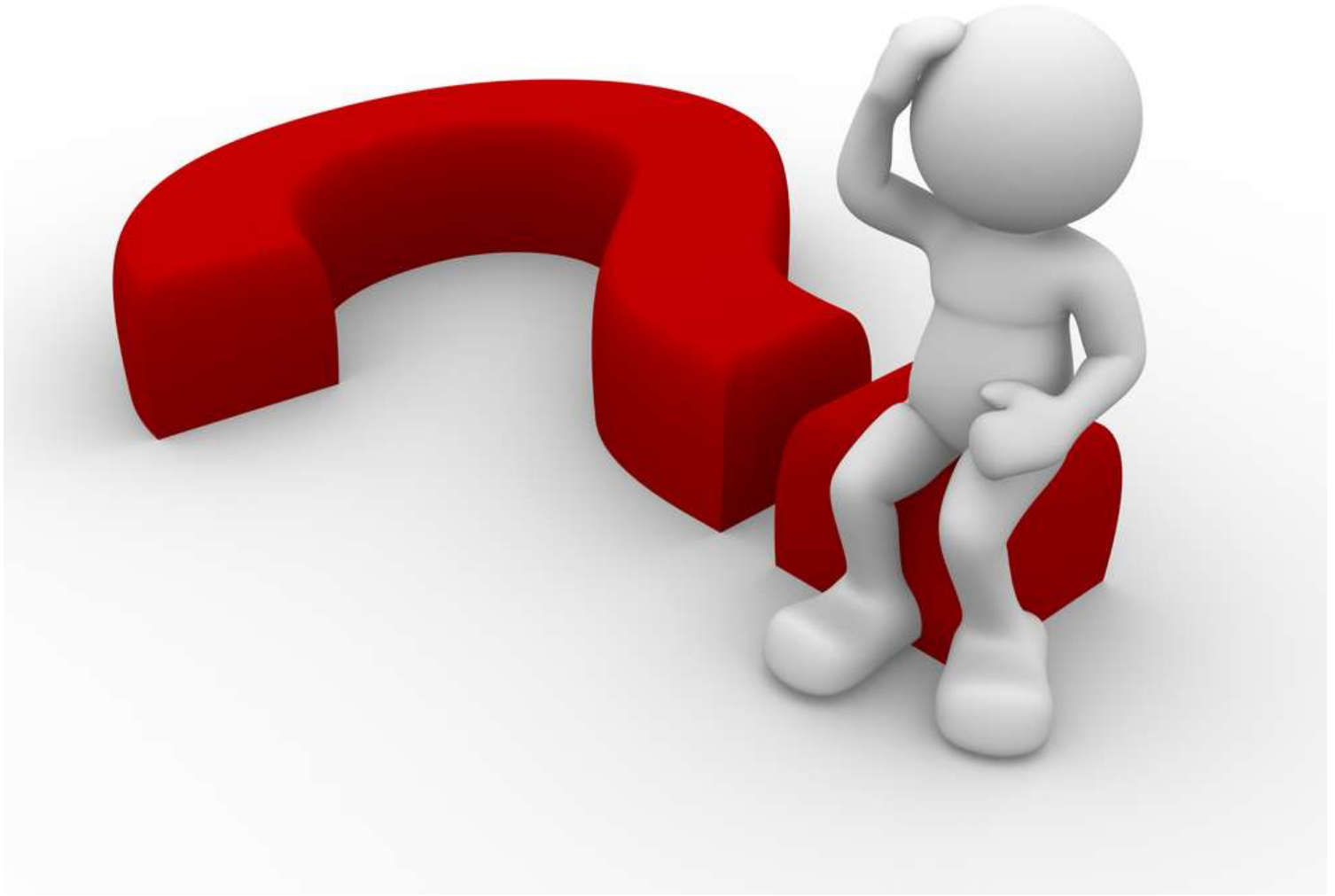
Ozone is designed to scale to tens of billions of files and blocks and, in the future, even more.

Small files or huge number of datanodes are no longer a limitation.

Honorable Mentions

- Apache Griffin (incubating): https://github.com/apache/griffin
- Apache Tika: https://tika.apache.org/
- Apache Metron: http://metron.apache.org/
- Apache Beam: https://projects.apache.org/project.html?beam

Questions



Next Meetup

- Docker on Hadoop
  - Feb 6th