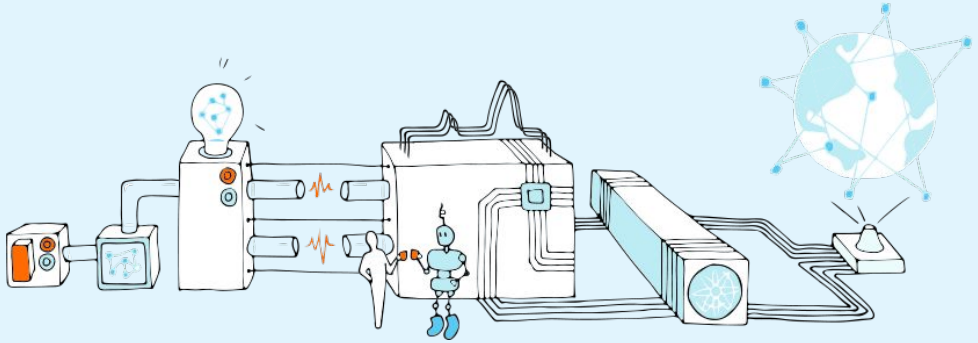
Hive and Presto for Big Data Analytics

Bang Dinh (bang.dinh@nfq.asia)

June 13, 2019



😎 Friendly & Social

Bang Dinh
dinhnhatbang

Backend Engineer
.NFQ AISA
HCMC, Viet Nam
bang.dinh@nfq.asia

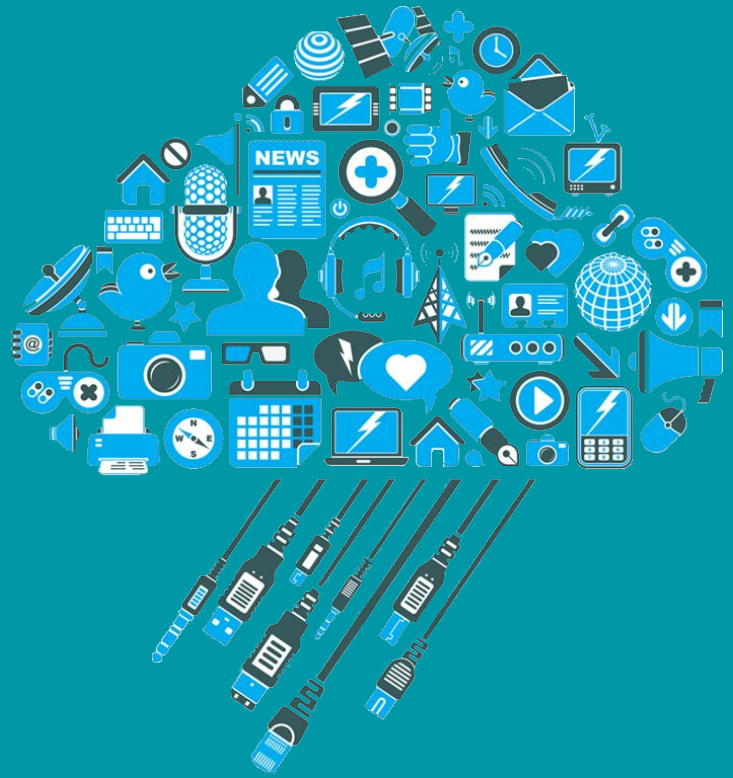
A little about me

- My name is Bang Dinh
- I was born in 1994
- Github: https://github.com/dinhnhatbang
- Linkedin: https://www.linkedin.com/in/bangdinh

Today's talk

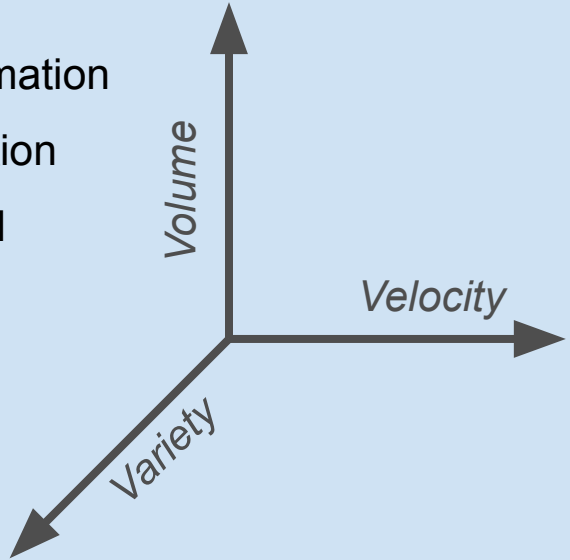
- ❏ What is big data?
- ❏ Big data technologies
- ❏ What is Presto?
- ❏ **Presto use cases**
- ❏ **Presto concepts**
- ❏ **Hadoop integration & Hive connector**
- ❏ **Demo**
- ❏ Q&A

What is big data?

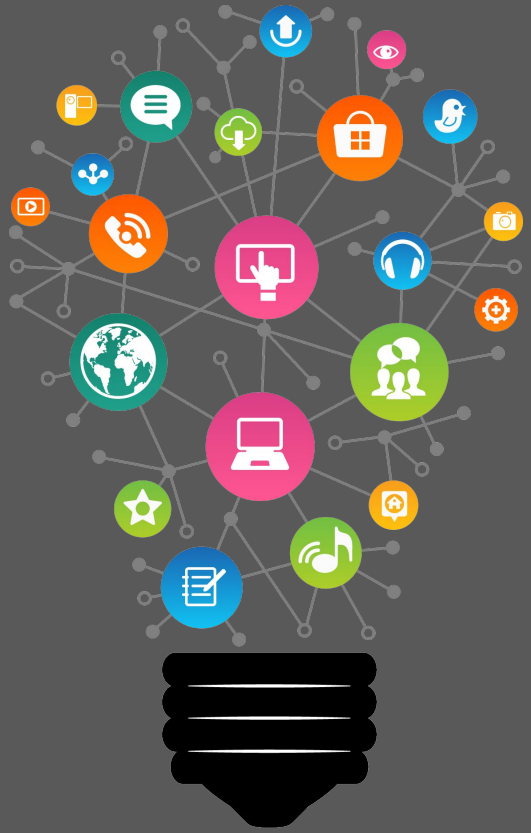


The Gartner's Big Data Definition

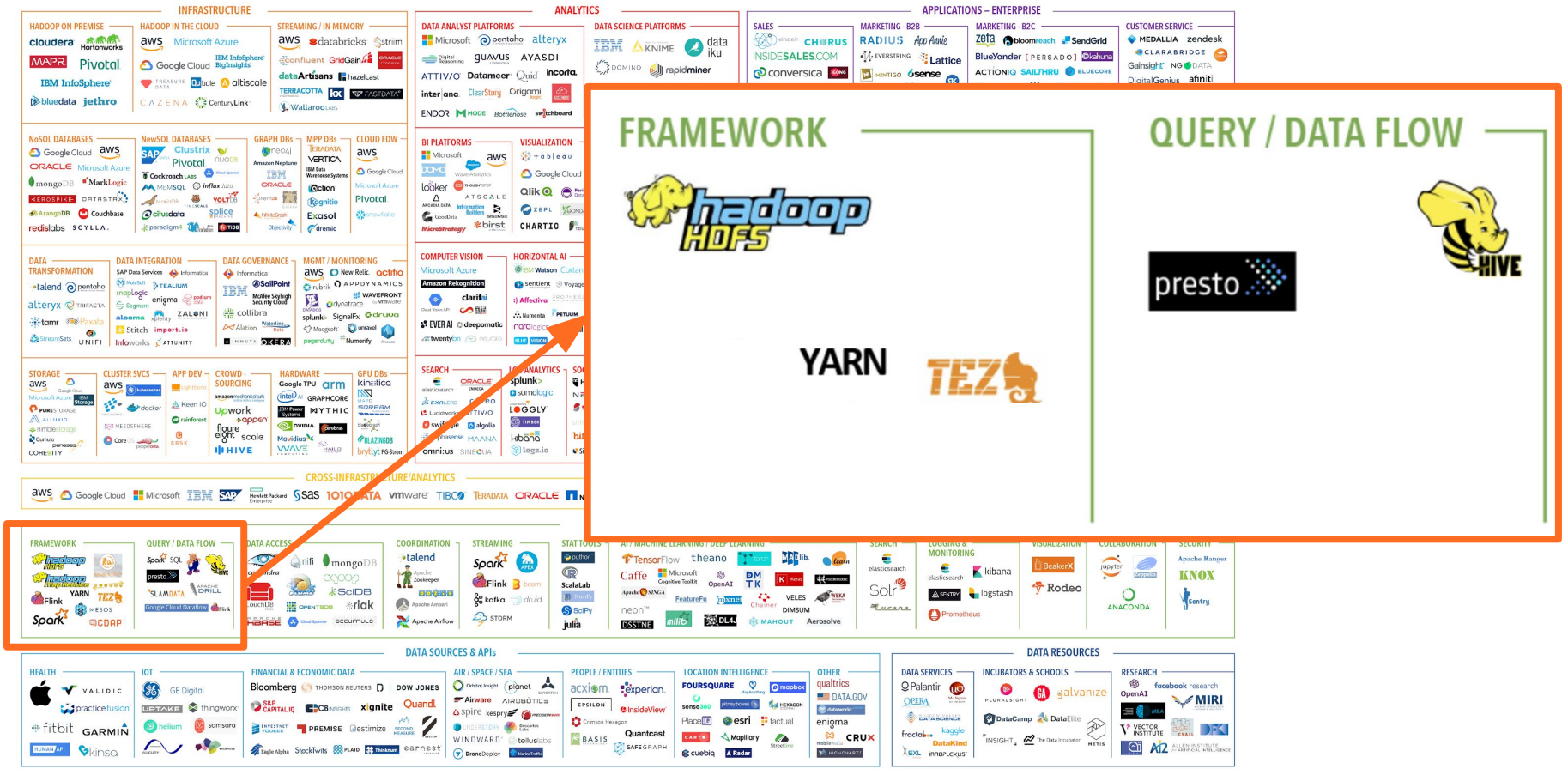
Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.



Big data technologies



BIG DATA & AI LANDSCAPE 2018



What is Presto or PrestoDB?

What is Presto or PrestoDB?

- Presto is a fast distributed SQL query engine for big data. Presto is suitable for interactive querying of petabytes of data.

Use cases

What Presto is not

- **Do not** mistake the fact that Presto understands SQL with it providing the features of a standard database.
- **Presto is not a general-purpose relational database.** It is not a replacement for databases like MySQL, PostgreSQL or Oracle.
- Presto was not designed to handle Online Transaction Processing (OLTP).

What Presto is

- Presto is a tool designed to efficiently query vast amounts of data using distributed queries. If you work with terabytes or petabytes of data, you are likely using tools that interact with Hadoop and HDFS.

- Presto was designed as an alternative to tools that query HDFS using pipelines of MapReduce jobs.

What Presto is

- Presto can be and has been extended to operate over different kinds of data sources.

- Presto was designed to handle data warehousing and analytics: data analysis, aggregating large amounts of data and producing reports. These workloads are often classified as Online Analytical Processing (OLAP).

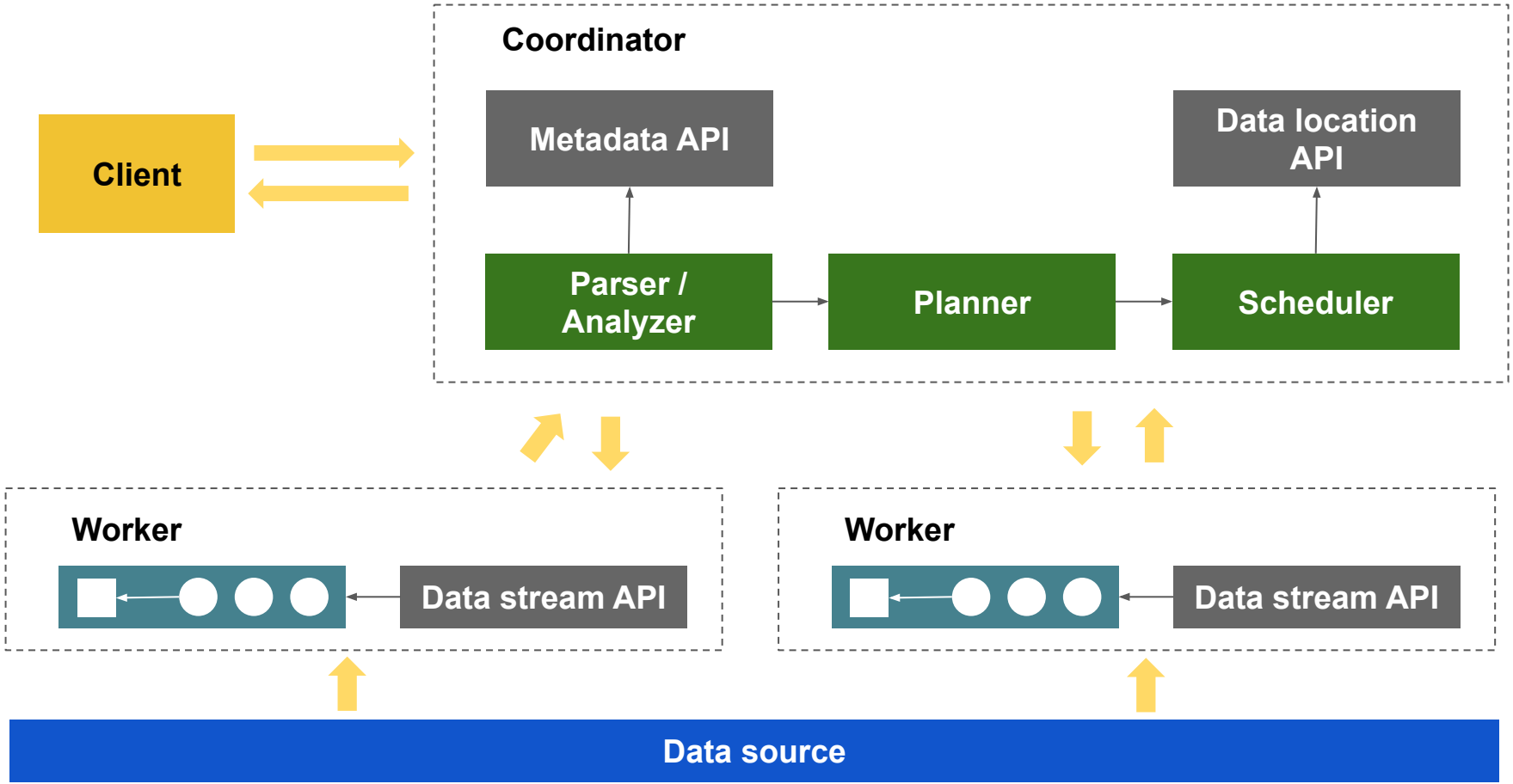
What is the difference between OLTP and OLAP?

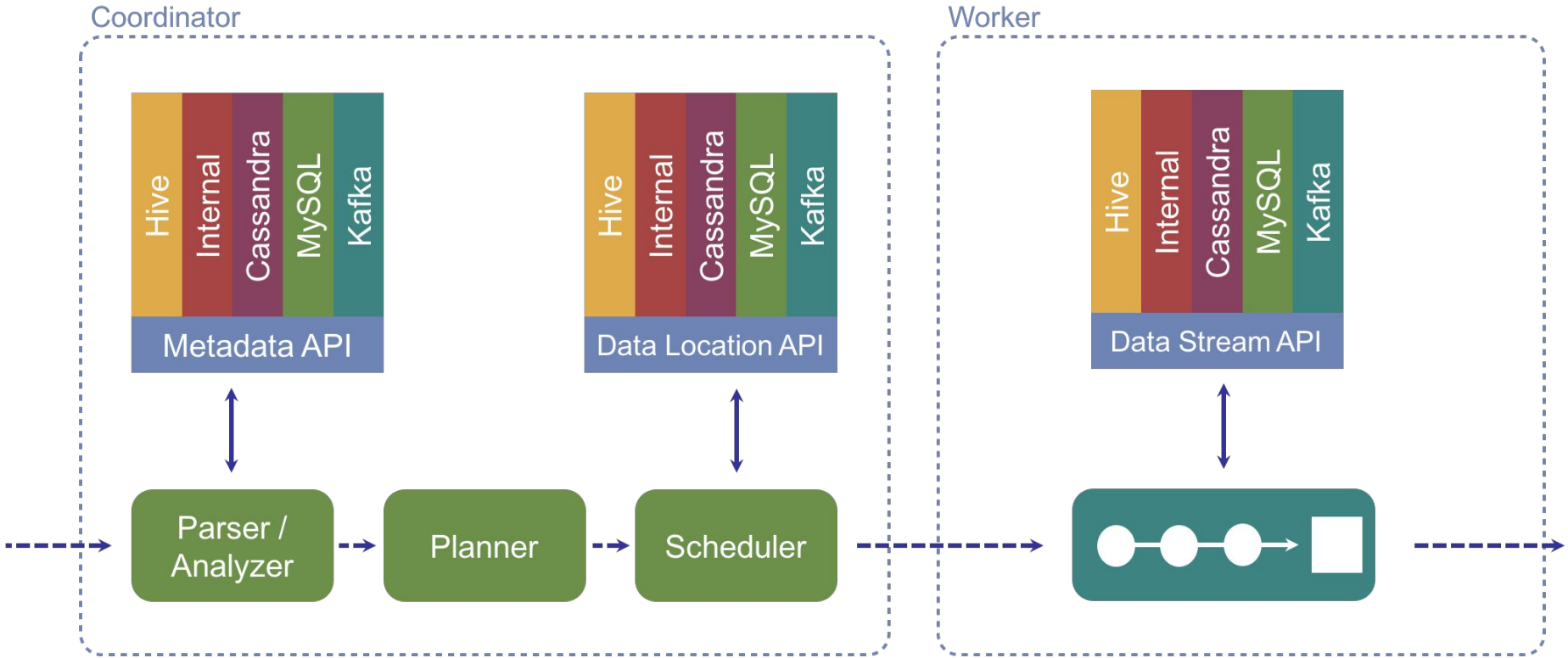
Who is using Presto?

Who is using Presto?

https://github.com/prestodb/presto/wiki/Presto-Users

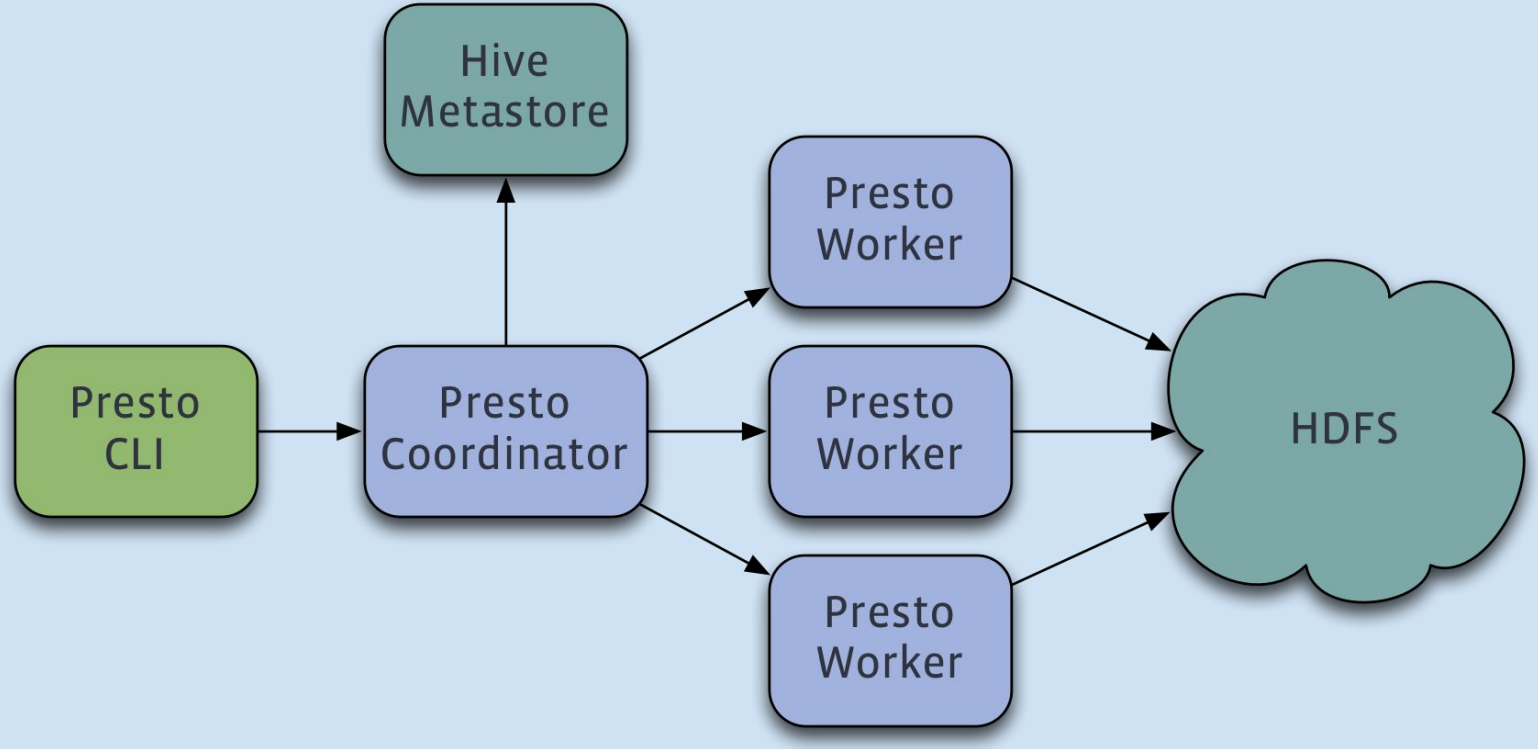
Concepts


Coordinator
Client
Metadata API
Data location API
Parser / Analyzer
Planner
Scheduler
Worker
Data stream API
Worker
Data stream API
Data source



Source: http://dharmeshkakadia.github.io/presto-hdinsight/

Hadoop integration & Hive connector


Hive
Metastore
Presto
Worker
Presto
CLI
Presto
Coordinator
Presto
Worker
HDFS
Presto
Worker

Hadoop Integration—Hive Connector

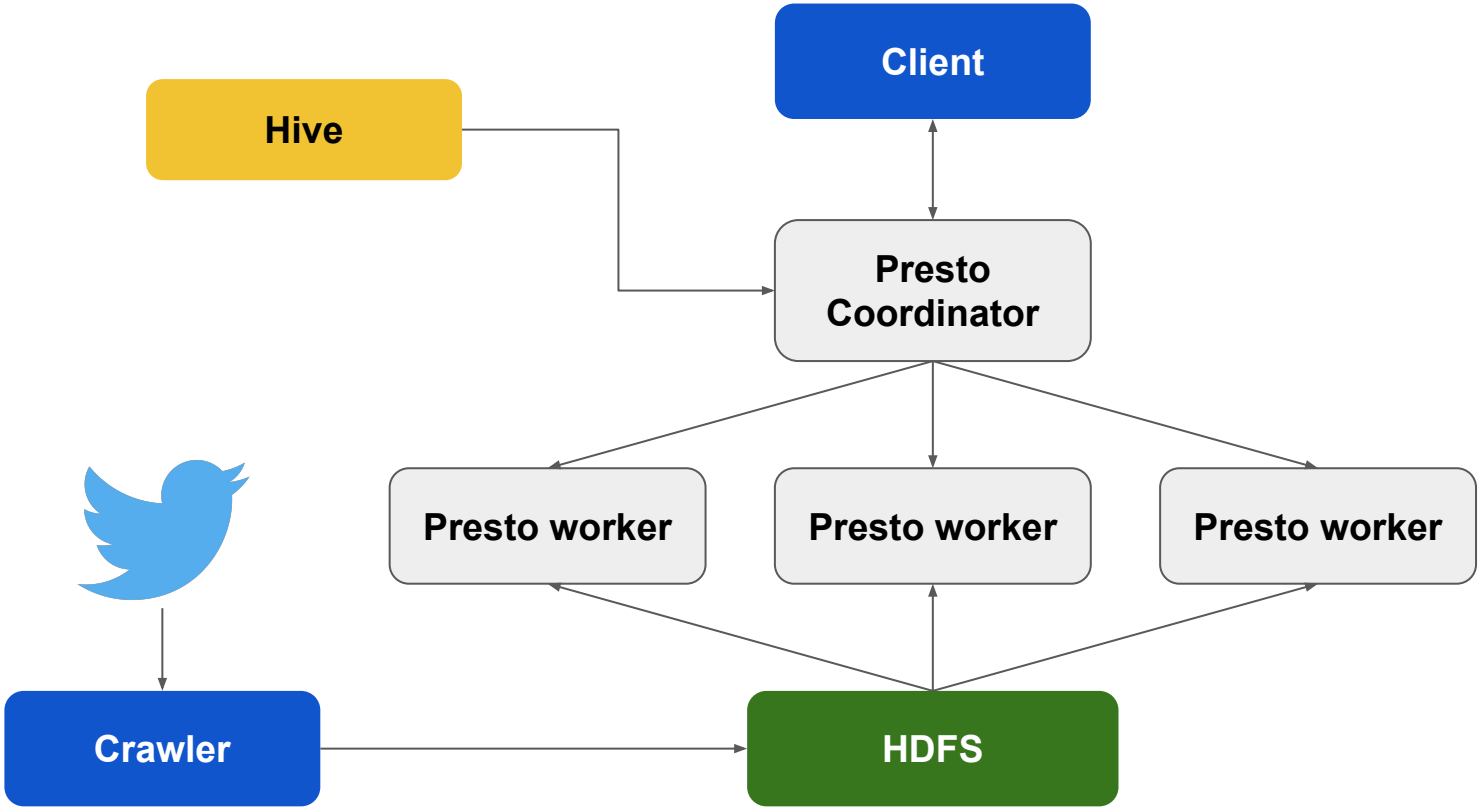
- The Hive connector allows querying data stored in HDFS and mapped with **Hive Metastore**.
- The supported file types are: ORC, Apache Parquet, Avro, RCFile, SequenceFile, JSON, Text.

Demo

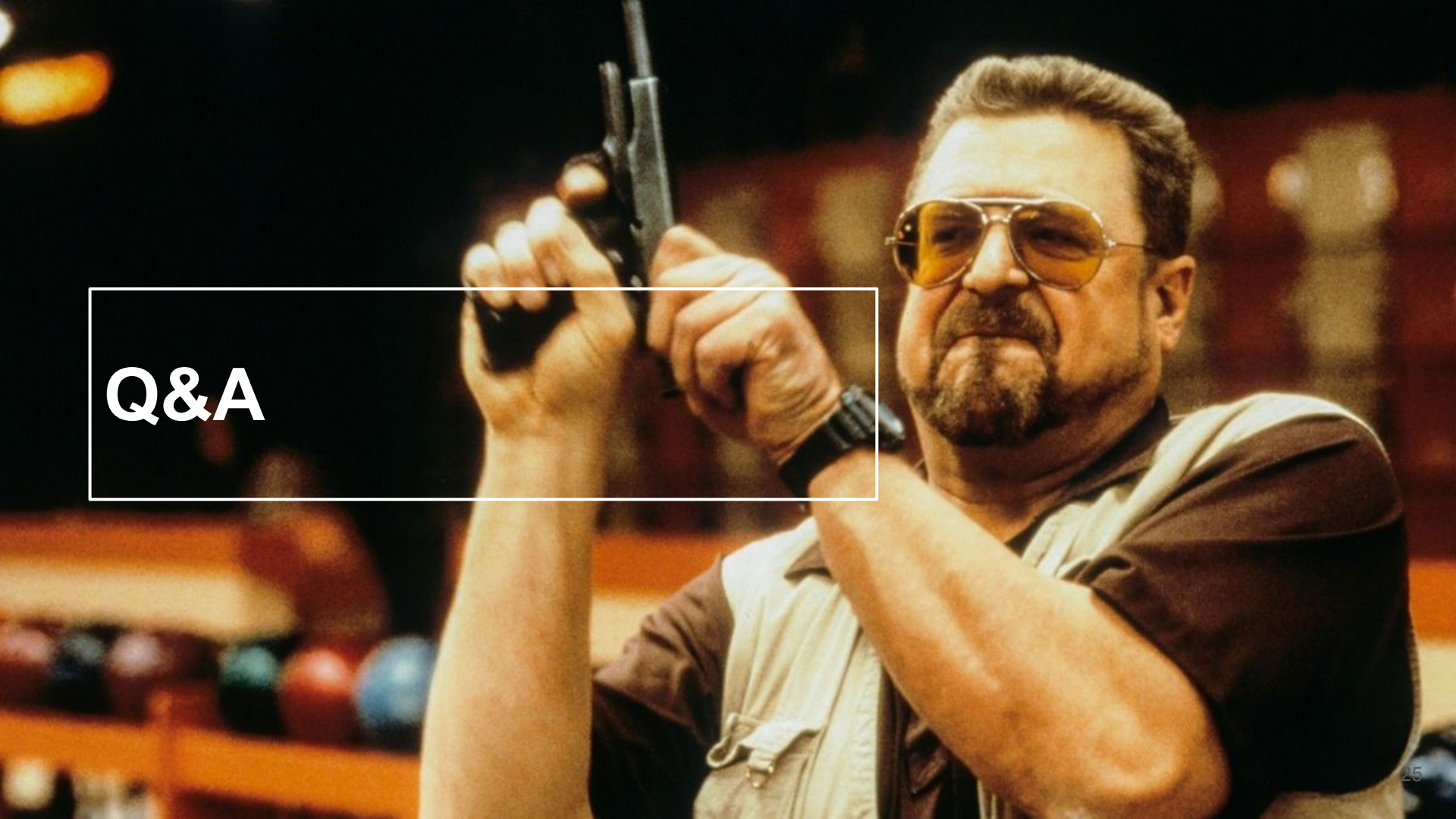
```
$ presto
presto:default> describe nation;
   Column    |  Type   | Null | Partition Key
-------------+---------+------+---------------
 n_nationkey | bigint  | true | false
 n_name      | varchar | true | false
 n_regionkey | bigint  | true | false
 n_comment   | varchar | true | false
(4 rows)

Query 20131105_005529_00080_ee7y3, FINISHED, 2 nodes
Splits: 2 total, 2 done (100.00%)
0:00 [8 rows, 446B] [23 rows/s, 1.29KB/s]

presto:default>
```



Github: https://github.com/dinhnhatbang/hive-presto-docker


Q&A

Thank you