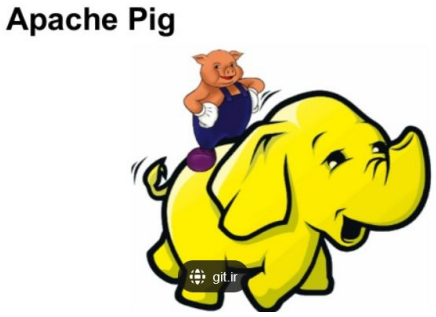
Hcatlog

Mahesh_16



Inside a Company

- Analysts use multiple tools for processing data.
  - Map Reduce, Hive, pig & more
- Analysts create many (derivative) datasets.
  - Different formats e.g. Csv, Json, Avro, Orc.
  - Files in HDFS .
  - Tables in Hive.

MR,Pig,Hive & Data Storage

- Hive reads data location , format and schema from metadata
  - Managed by Hive Metastore
- MapReduce encodes them in the application code.
- Pig specifies them in script
  - Schema can be provided by the Loader
- Conclusion
  - MapReduce and Pig are sensitive to metadata changes!

Office talk

- Jeff: Is your dataset already generated?
- Tom: Check in HDFS!
- Jeff: What is the location?
- Tom: /data/streams/20140101
- Jeff: hdfs dfs -ls /data/streams/20140101
- Not yet…. :(
- Tom: Check it later!

How to solve this problem ?

1) Location
2) Data schema
3) Format of files

HCatalog

- HCatalog is a table storage management tool for Hadoop, It exposes the tabular data of Hive metastore to other Hadoop applications. It enables users with different data processing tools (Pig, MapReduce) to easily write data onto a grid. It ensures that users don’t have to worry about where or in what format their data is stored.

- HCatalog works like a key component of Hive and it enables the users to store their data in any format and any structure.

HCatalog is mainly used for the following three reasons:

❑ Enabling right tool for right Job

- A workflow where data is loaded and normalized using MapReduce or Pig and then analyzed via Hive is very common.
- If all these tools share one metastore, then the users of each tool have immediate access to data created with another tool.
- No loading or transfer steps are required.

Capture processing states to enable sharing

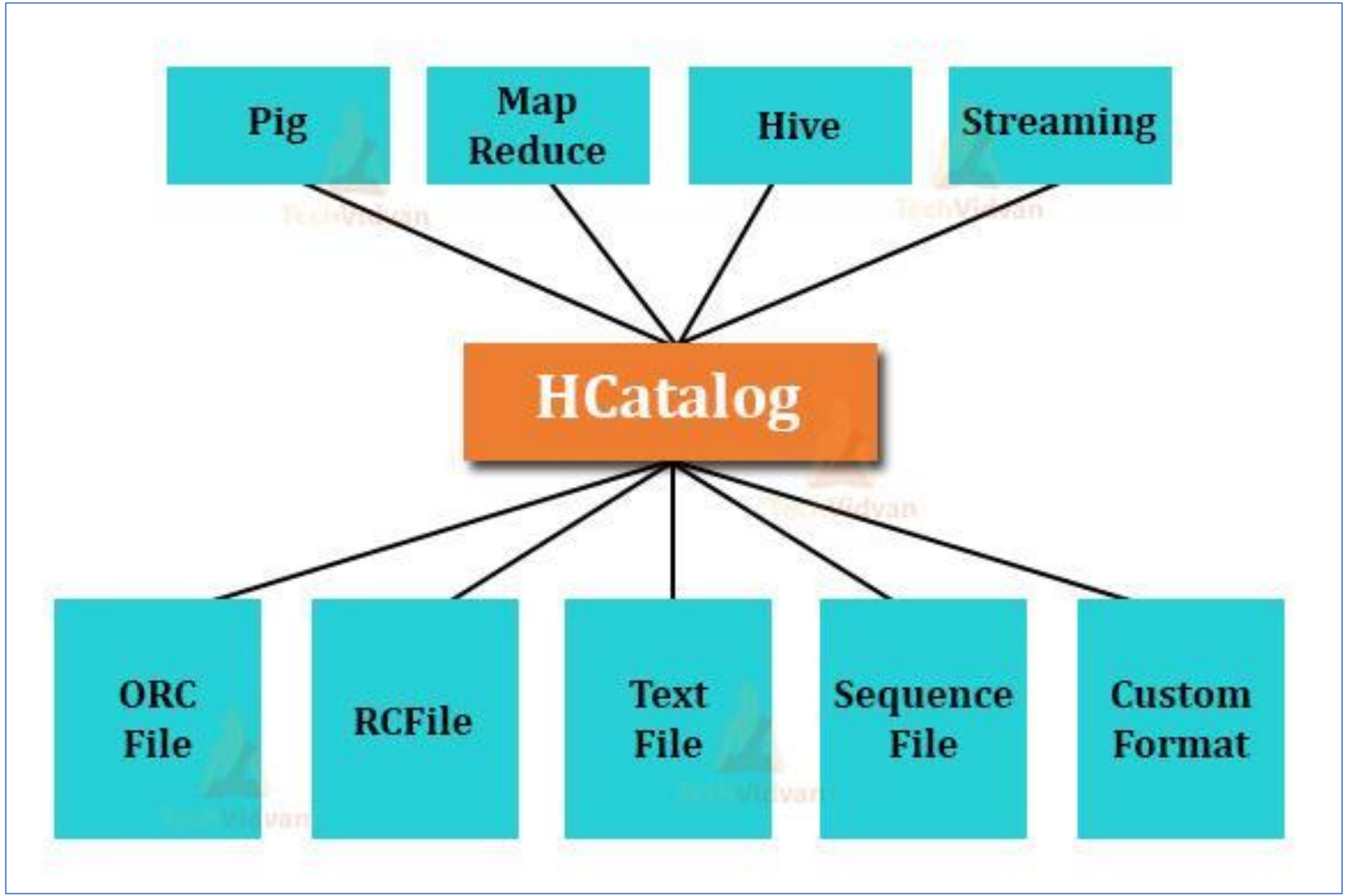
- HCatalog can publish your analytics results. So the other programmer can access your analytics platform via “REST”.
- The schemas which are published by you are also useful to other data scientists.
- The other data scientists use your discoveries as inputs into a subsequent discovery.

Integrating Hadoop with everything

- Apache Hadoop, because of its reliable storage and distributed processing, opens up a lot of opportunities for the businesses. However, in order to increase the adoption of Hadoop, it must work with existing tools.
- Organizations want to enjoy the value of Hadoop without learning an entirely new toolset.
- The REST services open up the platform to the businesses with the familiar API and the SQL-like language. The Enterprise data management system uses HCatalog to integrate more deeply with the Apache Hadoop platform..

HCatalog Architecture

- HCatalog is built on the top of Hive metastore and incorporates Hive DDL commands.
- It provides read and write interfaces for MapReduce and Pig


Pig
Map Reduce
Hive
Streaming
HCatalog
ORC File
RCFile
Text File
Sequence File
Custom Format

Custom Formats

- **A custom format can be supported**
  - But `InputFormat`, `OutputFormat`, and Hive `SerDe` must be provided

Pig Interface - HCatLoader

- **Consists of `HCatLoader` and `HCatStorer`**
- **`HCatLoader` read data from a dataset**
  - Indicate which partitions to scan by following the `load` statement with a partition filter statement

```
raw = load 'streams' using HCatLoader();
valid = filter raw by date = '20140101' and isValid(duration);
```

Pig Interface - HCatStorer

- **HCatStorer writes data to a dataset**
  - A specification of partition keys can be also provided
  - Possible to write to a single partition or multiple partitions

```
store valid into 'streams_valid' using HCatStorer
('date=20110924');
```

MapReduce Interface

- **Consists of `HCatInputFormat` and `HCatOutputFormat`**
- **`HCatInputFormat` accepts a dataset to read data from**
  - Optionally, indicate which partitions to scan
- **`HCatOutputFormat` accepts a dataset to write to**
  - Optionally, indicated with partition to write to
  - Possible to write to a single partition or multiple partitions

Hive Interface

- **There is no Hive-specific interface**
  - Hive can read information from HCatalog directly
- **Actually, HCatalog is now a submodule of Hive**

Conclusion

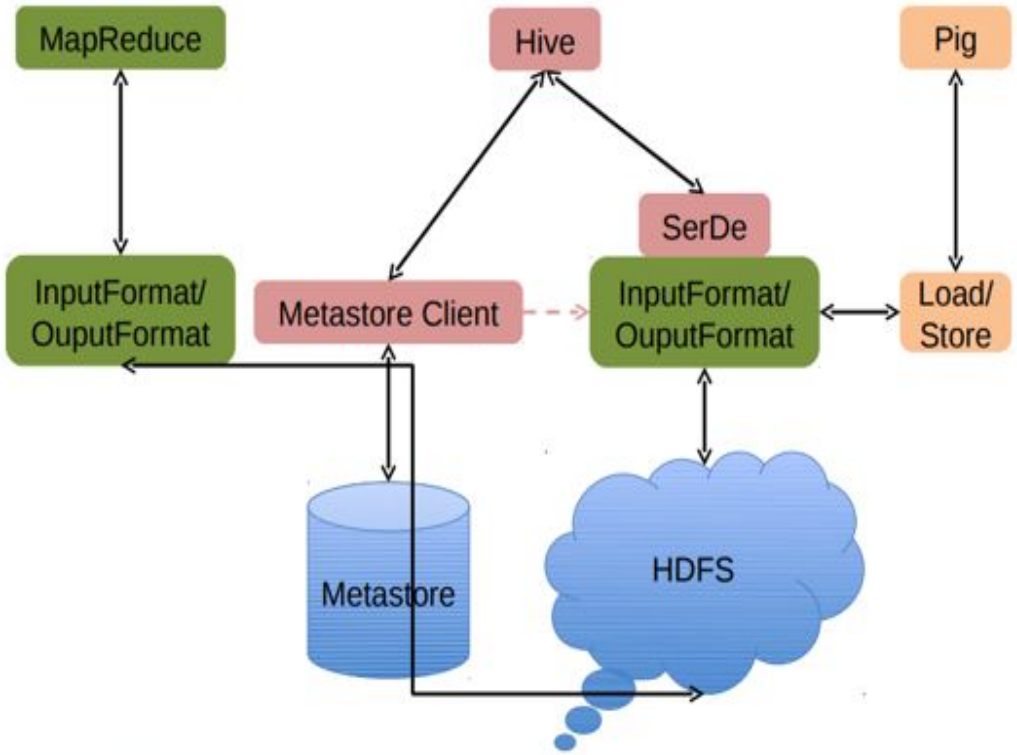
- **HCatalog enables non-Hive projects to access Hive tables**

HCatalog Features

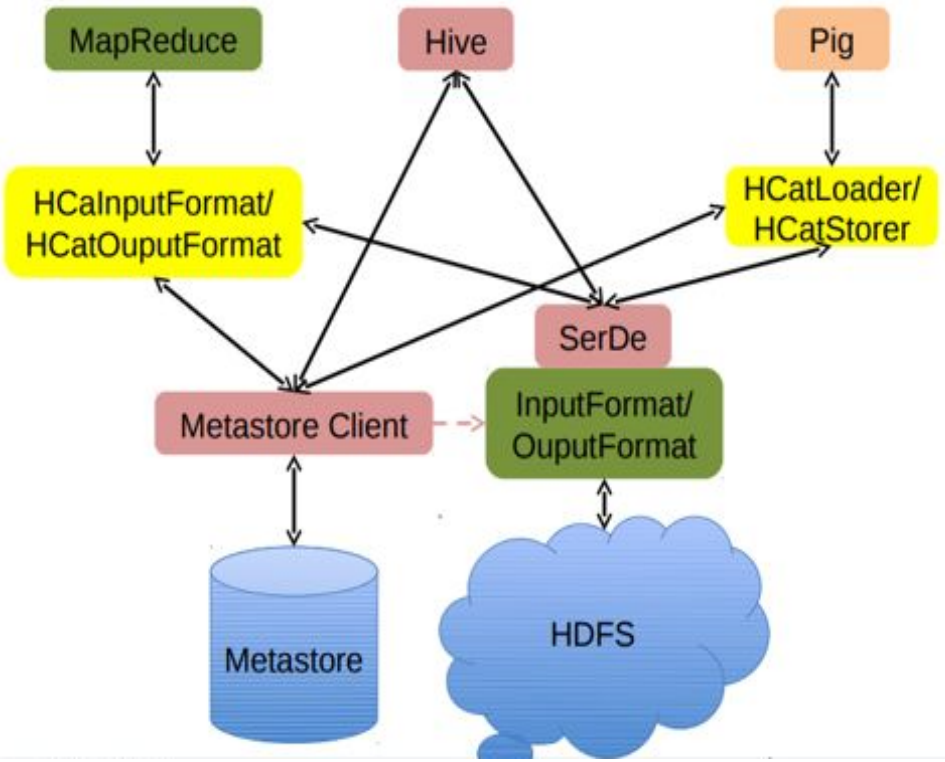
1) It assists the integration with the other tools and provides read and write interfaces for MapReduce, Pig, and Hive.
2) HCatalog provides the shared schema and data types for the Hadoop tools. So we don’t have to type the data structures in each program explicitly.
3) HCatalog exposes the information as a Rest Interface for the external data access.
4) HCatalog provides APIs and the web service wrapper for accessing the metadata in the Hive metastore.

Example

Hadoop Ecosystem



Opening up Metadata to MR & Pig



Top Companies Using Apache Hadoop Technology

1) A9.com
2) Adobe
3) Alibaba
4) AOL
5) ARA.COM.TR
6) CRS4
7) eBay
8) eCircle
9) Facebook
10) Spotify

Summary

- In short, we can say that HCatalog is the table storage management tool for Apache Hadoop, which exposes the tabular data of Apache Hive metastore to the other Hadoop applications.
- It provides read and write interfaces for MapReduce, Hive, and Pig.
- It ensures users not to worry about the location and the format in which their data is stored.

THANK YOU