Apache Sqoop, Apache Pig & Apache Hive

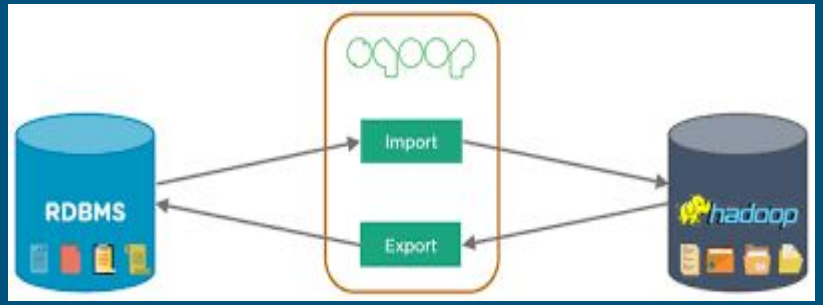
Apache Sqoop

Apache SQOOP (SQL-to-Hadoop) is tool designed to support bulk export and import of data into HDFS from structured data store such as relational databases, enterprise data warehouses, and NoSQL systems. It is a data migration tool based upon a connector architecture which supports plugins to provide connectivity to new external system.

Apache Sqoop is a tool that is extensively used to transfer large amounts of data from Hadoop to the relational database servers and vice-versa. Sqoop can be used to import the various types of data from Oracle, MySQL, and other databases.

Apache Sqoop

Analytical processing using Hadoop requires loading of huge amounts of data from diverse sources into Hadoop clusters. This process of bulk data load into Hadoop, from heterogeneous sources and then processing it, comes with a certain set of challenges. Maintaining and ensuring data consistency and ensuring efficient utilization of resources, are some factors to consider before selecting the right approach for data load.



Apache Sqoop

Sqoop supports data imported into the following services:

- HDFS
- Hive
- HBase –HBase is a non-relational database. It is a columnar and provides fault-tolerant storage and quick access to large quantities of sparse data. It also adds transactional capabilities to Hadoop. Allowing users to conduct updates, inserts and deletes.
- Hcatalog
- Accumulo

Features of Sqoop

Full Load

Data loading directly to HIVE

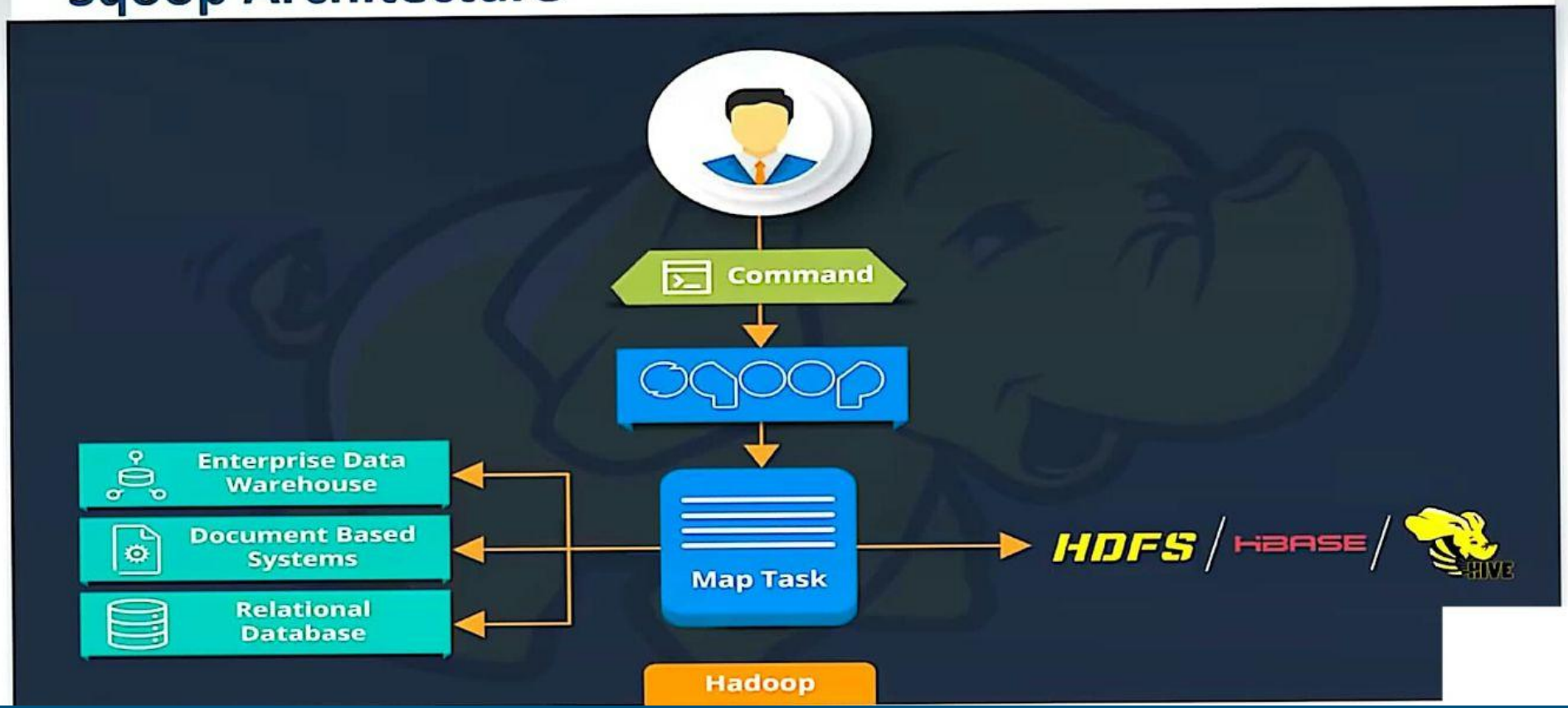
Incremental Load

Kerberos Security Integration

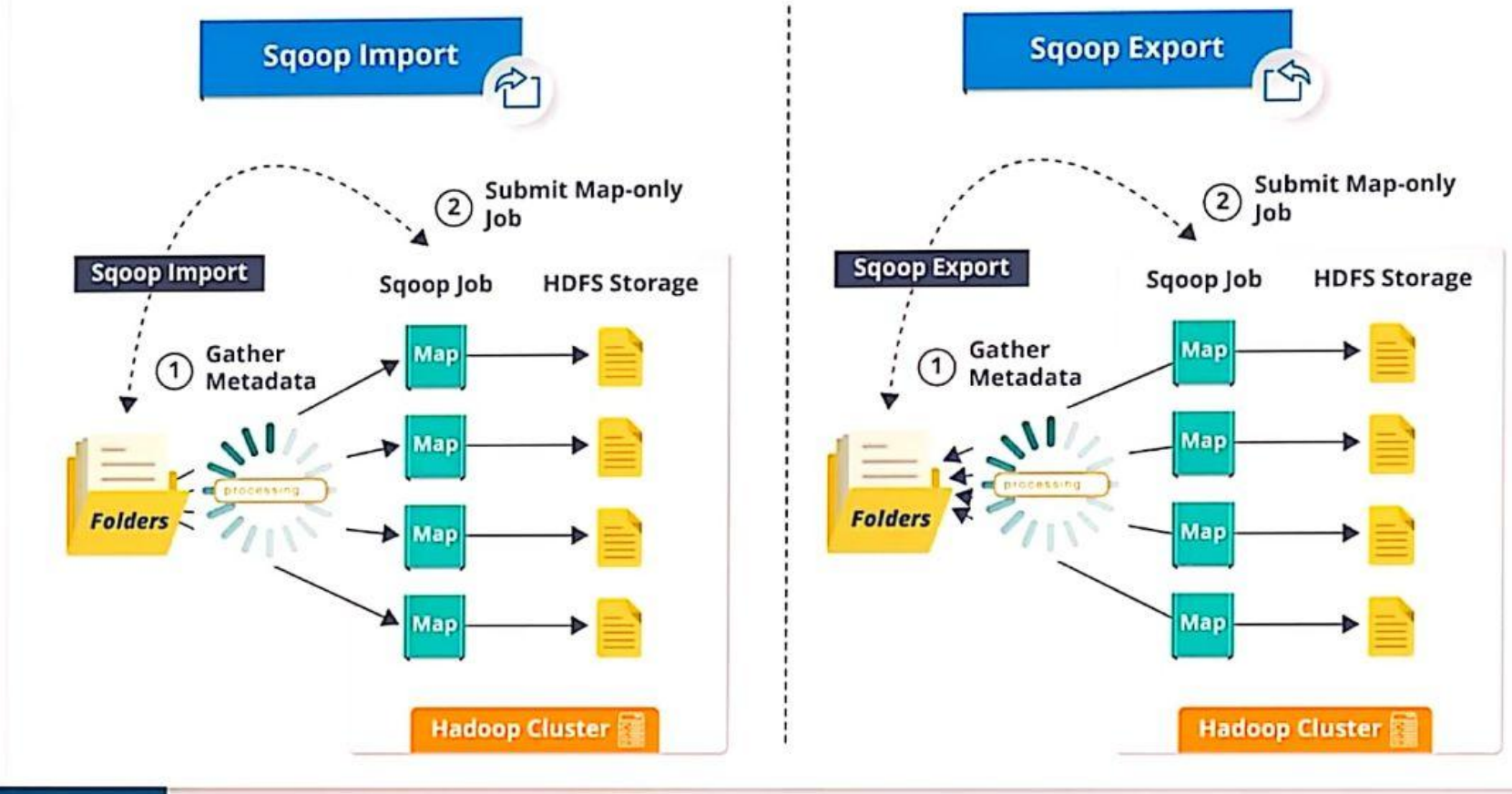
Parallel Import/Export

Compression

Sqoop Architecture



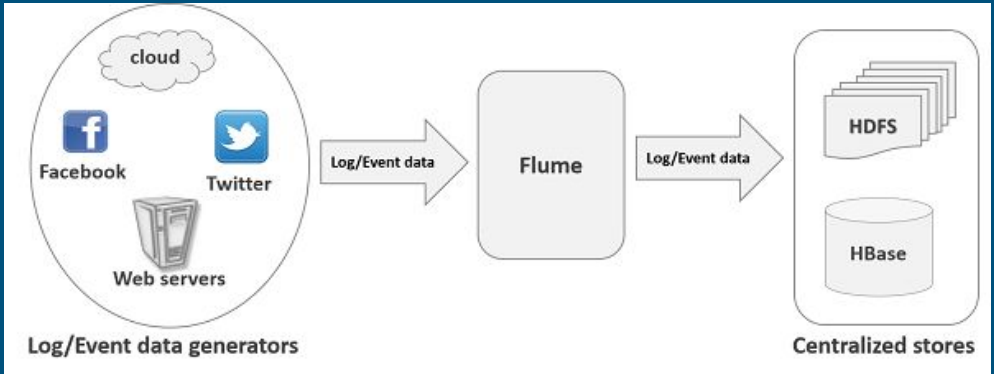
How Sqoop Import & Export Works?



Apache Flume

- Apache Flume is a tool for data ingestion in HDFS. It collects, aggregates and transports large amount of streaming data such as log files, events from various sources like network traffic, social media, email messages etc. to HDFS. Flume is a highly reliable & distributed.
- The main idea behind the Flume’s design is to capture streaming data from various web servers to HDFS. It has simple and flexible architecture based on streaming data flows. It is fault-tolerant and provides reliability mechanism for Fault tolerance & failure recovery.

Apache Flume: Architecture



The flume agent has 3 components: source, sink and channel.
Source: It accepts the data from the incoming streamline and stores the data in the channel.
Channel: In general, the reading speed is faster than the writing speed. Thus, we need some buffer to match the read & write speed difference. Basically, the buffer acts as a intermediary storage that stores the data being transferred temporarily and therefore prevents data loss. Similarly, channel acts as the local storage or a temporary storage between the source of data and persistent data in the HDFS.
Sink: Then, our last component i.e. Sink, collects the data from the channel and commits or writes the data in the HDFS permanently.

Apache Sqoop , Flume and HDFS

Sqoop	Flume	HDFS
Sqoop is used for importing data from structured data sources such as RDBMS.	Flume is used for moving bulk streaming data into HDFS.	HDFS is a distributed file system used by Hadoop ecosystem to store data.
Sqoop has a connector based architecture. Connectors know how to connect to the respective data source and fetch the data.	Flume has an agent-based architecture. Here, a code is written (which is called as 'agent') which takes care of fetching data.	HDFS has a distributed architecture where data is distributed across multiple data nodes.
HDFS is a destination for data import using Sqoop.	Data flows to HDFS through zero or more channels.	HDFS is an ultimate destination for data storage.
Sqoop data load is not event-driven.	Flume data load can be driven by an event.	HDFS just stores data provided to it by whatsoever means.

Apache Pig

Pig is an abstraction over MapReduce. Pig runs on Hadoop. So, it makes use of both the Hadoop Distributed File System (HDFS) and Hadoop's processing system, MapReduce. Data flows are executed by an engine. It is used to analyze data sets as data flows. It includes a high-level language called Pig Latin for expressing these data flows.

Apache Pig was designed by Yahoo as it is easy to learn and work with. So, Pig makes Hadoop quite easy. Apache Pig was developed because MapReduce programming was getting quite difficult and many MapReduce users are not comfortable with declarative languages. Now, Pig is an open-source project under Apache.

Apache Pig

Features of Apache Pig:

- For performing several operations Apache Pig provides rich sets of operators like the filters, join, sort, etc.
- Easy to learn, read and write. Especially for SQL-programmer, Apache Pig is a boon.
- Apache Pig is extensible so that you can make your own user-defined functions and process.
- Join operation is easy in Apache Pig.
- Fewer lines of code.
- Apache Pig allows splits in the pipeline.
- The data structure is multivalued, nested, and richer.
- Pig can handle the analysis of both structured and unstructured data.

Apache Pig

Applications of Apache Pig:

- For exploring large datasets Pig Scripting is used.
- Provides the supports across large data-sets for Ad-hoc queries.
- In the prototyping of large data-sets processing algorithms.
- Required to process the time sensitive data loads.
- For collecting large amounts of datasets in form of search logs and web crawls.
- Used where the analytical insights are needed using the sampling.

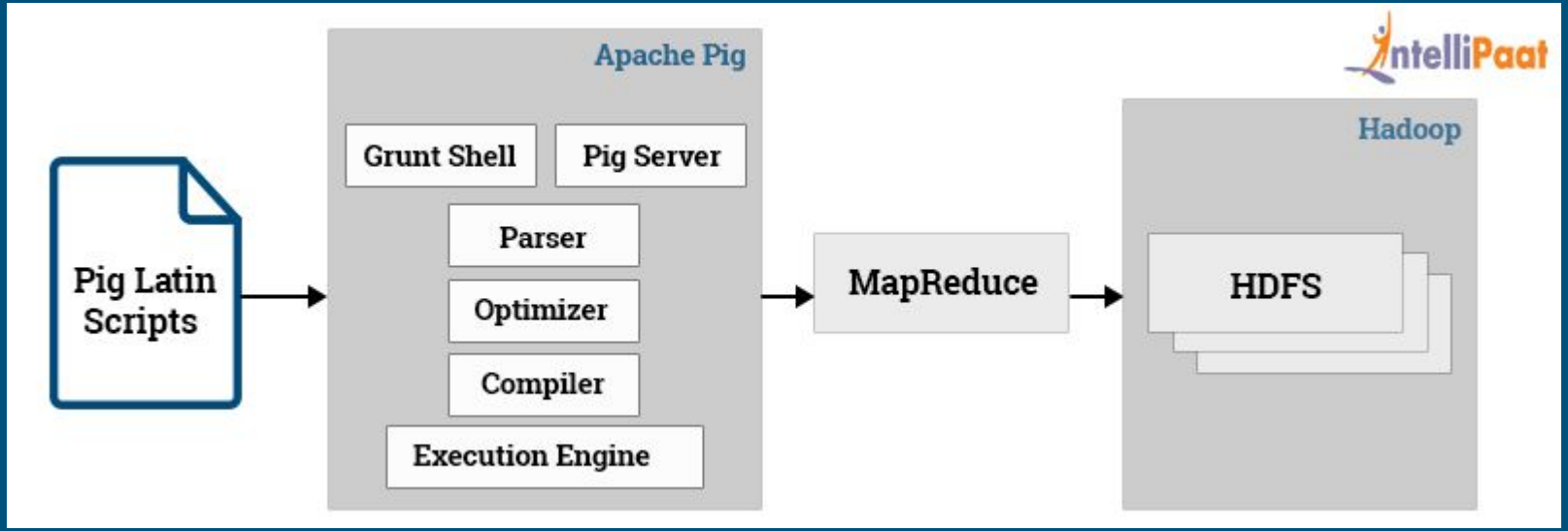
Apache Sqoop Vs Pig

Pig	Sqoop
Apache Pig is a tool for analytics which is used to analyze data stored in HDFS.	Apache Sqoop is a tool to importing structured data from RDBMS to HDFS or exporting data from HDFS to RDBMS.
We can import the data from Sql databases into hive rather than NoSql Databases.	It can integrate with any external data sources with HDFS i.e Sql , NoSql and Data warehouses as well using this tool at the same time we export it as well since this can be used as bi-directional ways

Apache Sqoop Vs Pig cont....

Pig can be used for following purposes ETL data pipeline, Research on raw data.	Important Sqoop control commands to import RDBMS data are Append, Columns and Where
The pig Metastore stores all info about the tables. And we can execute spark sql queries because spark can interact with pig Metastore.	Sqoop metastore is a tool for using hosts in a shared metadata repository. Multiple users and remote users can define and execute saved jobs defined in metastore.

Apache Pig


Apache Pig
IntelliPaat
Hadoop
Grunt Shell
Pig Server
Parser
Optimizer
Compiler
Execution Engine
Pig Latin Scripts
MapReduce
HDFS

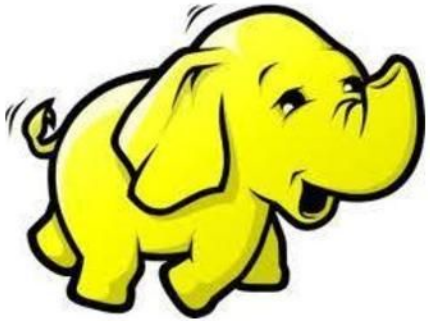
- Diagnostic Operators
- Grouping & Joining
- Combining & Splitting
- Filtering
- Sorting

TechVidvan

Apache Pig

Operators

PIG vs MapReduce



MapReduce

- Powerful model for parallelism.
- Based on a rigid procedural structure.
- Provides a good opportunity to parallelize algorithm.
- Must think in terms of map and reduce functions
- More than likely will require Java programmers

PIG

- Higher level procedural data flow language
- Similar to SQL query where the user specifies the what and leaves the “how” to the underlying processing engine.

Apache Hive

The Apache Hive is a data warehouse software that lets you read, write and manage huge volumes of datasets that are stored in a distributed environment using SQL. It is possible to project structure onto data that is in storage. Users can connect to Hive using a JDBC driver and a command-line tool.

Hive is an open system. We can use Hive for analyzing and querying in large datasets of Hadoop files. It's similar to SQL. The present version of Hive is 0.13.1.

Apache Hive

Hive supports ACID transactions: The full form of ACID is Atomicity, Consistency, Isolation, and Durability. ACID transactions are provided at the row levels, there are Insert, Delete, and Update options so that Hive supports ACID transactions.

Hive is not considered a full database. The design rules and regulations of Hadoop and HDFS put restrictions on what Hive can do.

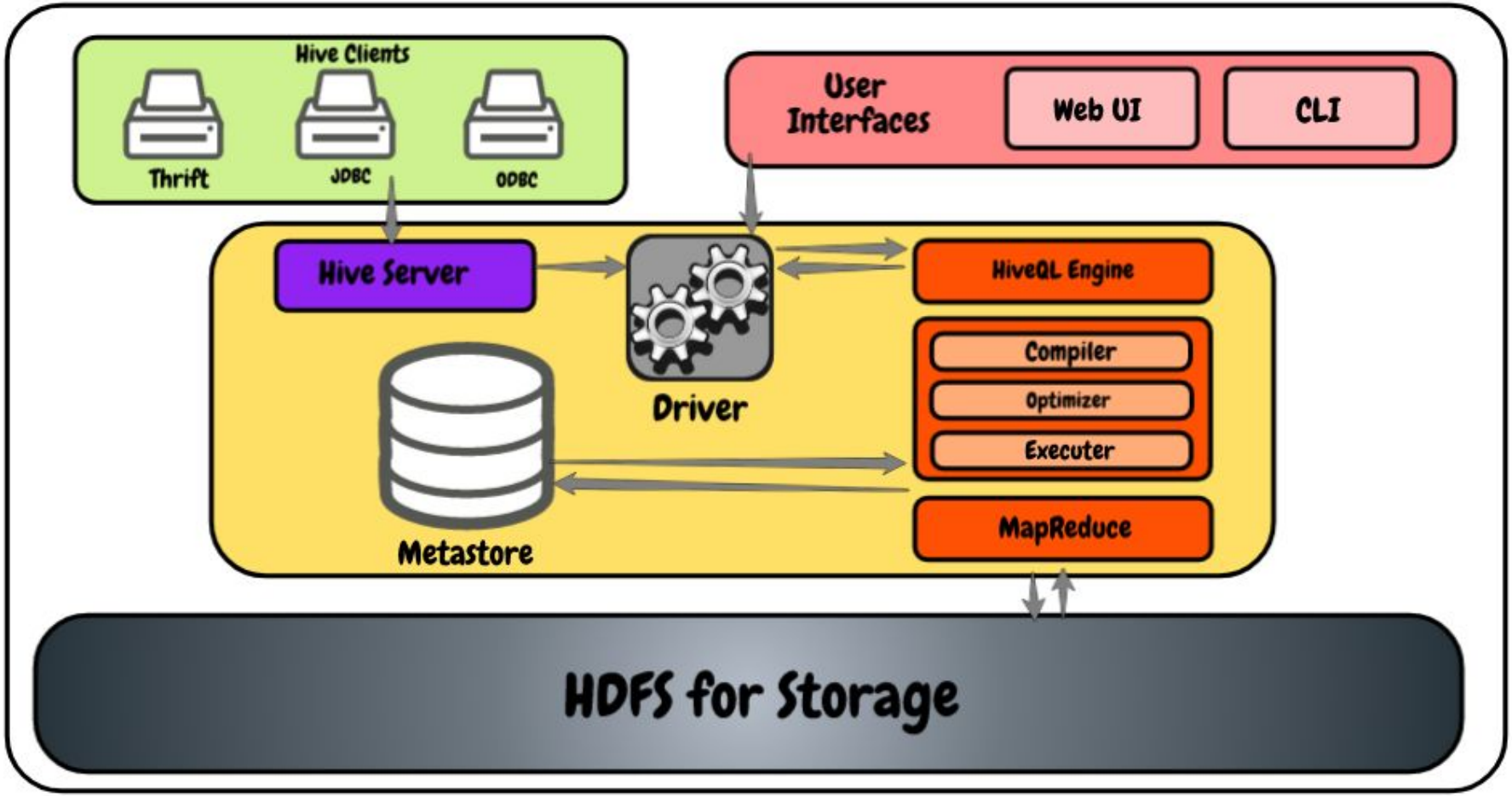
Apache Hive

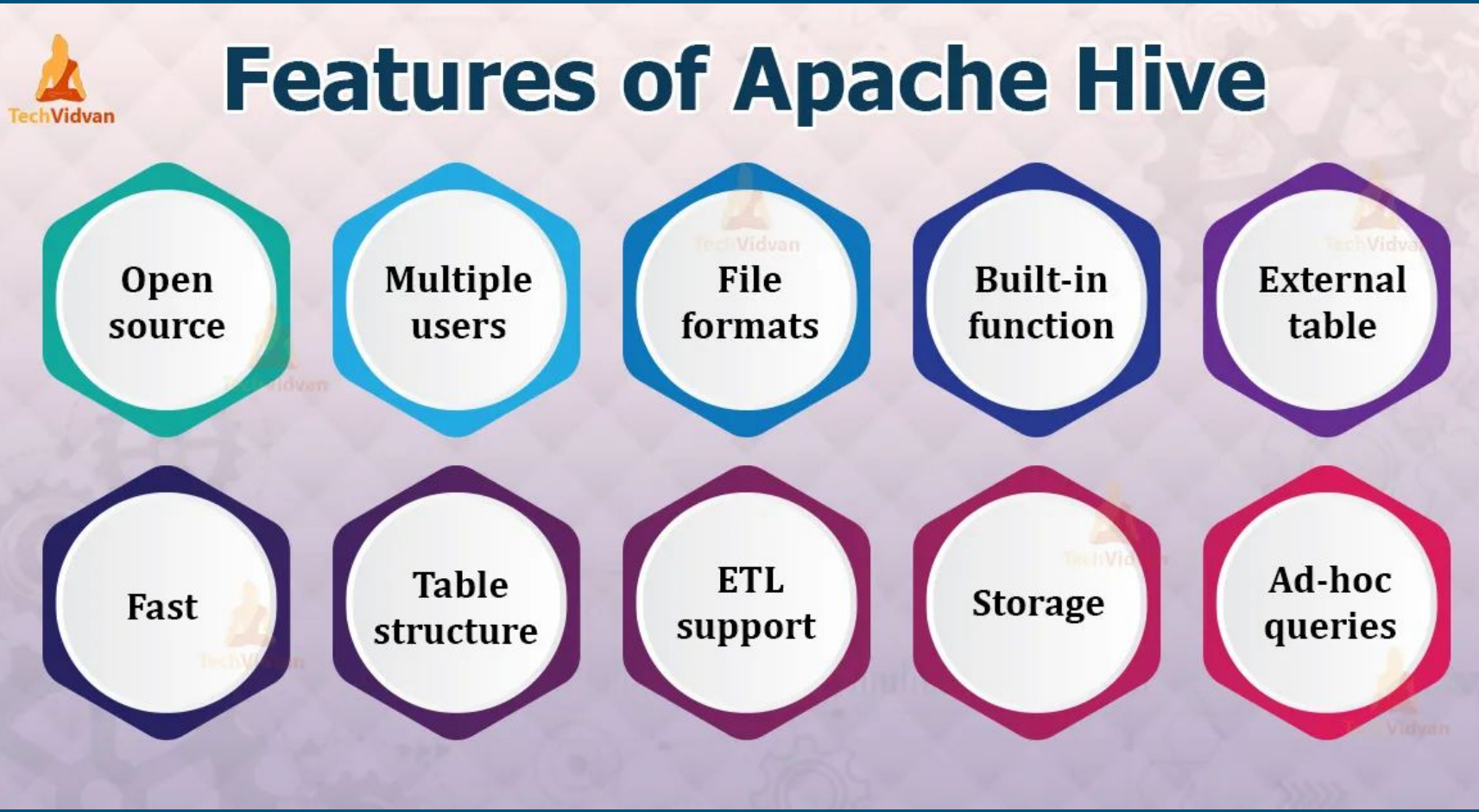
data warehouse applications

- Analyzing the relatively static data
- Less Responsive time
- No rapid changes in data.

Hive doesn't provide fundamental features required for OLTP, Online Transaction Processing. Hive is suitable for data warehouse applications in large data sets.

Hive Architecture




TechVidvan
Features of Apache Hive
Open source
Multiple users
File formats
Built-in function
External table
Fast
Table structure
ETL support
Storage
Ad-hoc queries

Pig	Hive
Procedural Data Flow Language	Declarative SQLish Language
For Programming	For creating reports
Mainly used by Researchers and Programmers	Mainly used by Data Analysts
Operates on the client side of a cluster.	Operates on the server side of a cluster.
Does not have a dedicated metadata database.	Makes use of exact variation of dedicated SQL DDL language by defining tables beforehand.
Pig is SQL like but varies to a great extent.	Directly leverages SQL and is easy to learn for database experts.
Pig supports Avro file format.	Hive does not support it.

Thank you