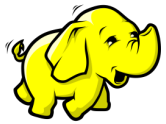
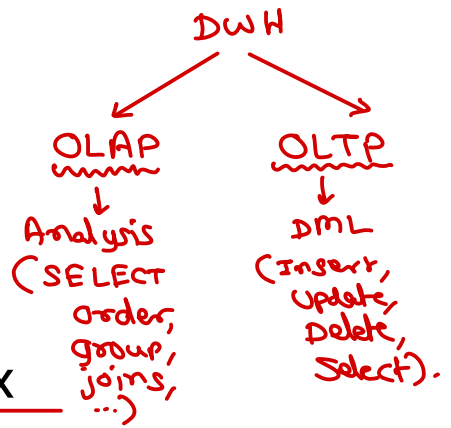


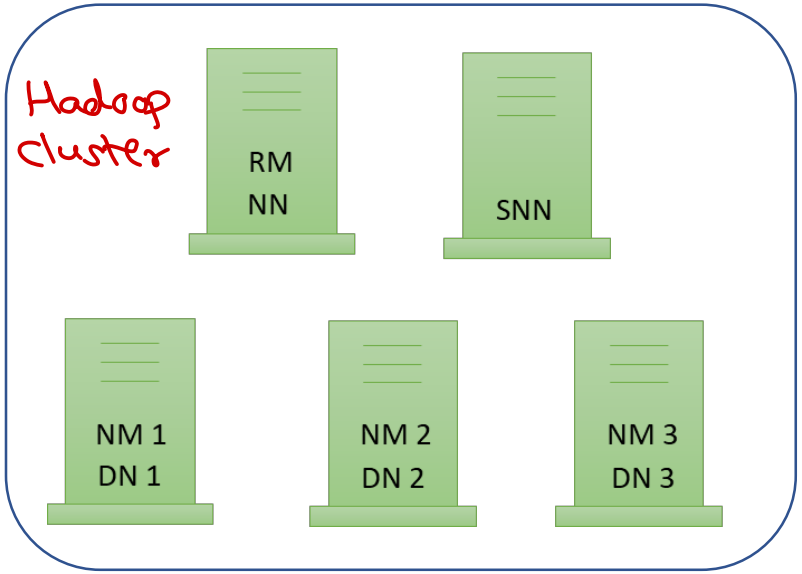
Apache Hive

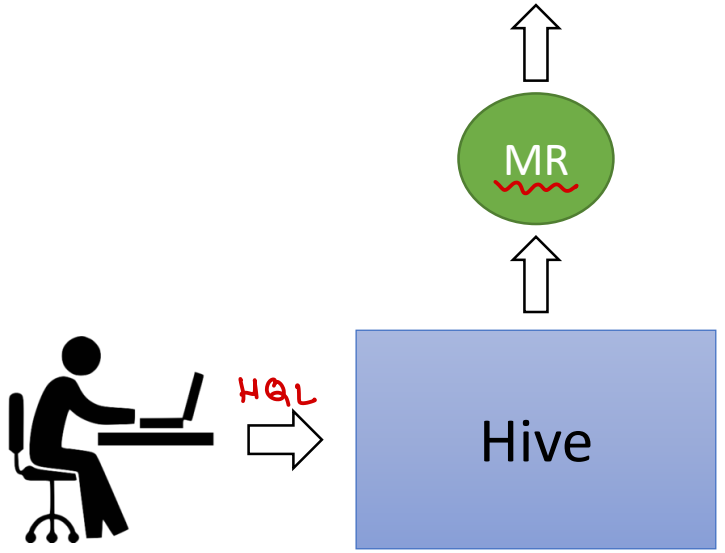
Sunbeam Infotech

Hive Introduction

- History
  - Facebook data ingestion into Hadoop
    - 10s GB/day – 2006
    - 1 TB/day – 2007
    - MySQL/Oracle database limitations
  - Processing Hadoop data using MR is complex
  - Developed Hive to convert SQL queries into MR
  - Open sourced under Apache license (2010)
- Hive is client software that convert Hive QL queries to MR.
- Hive QL is similar to SQL with many extended features.
- Hive manage structured data.
- Hive is data warehouse (OLAP) built for Hadoop.

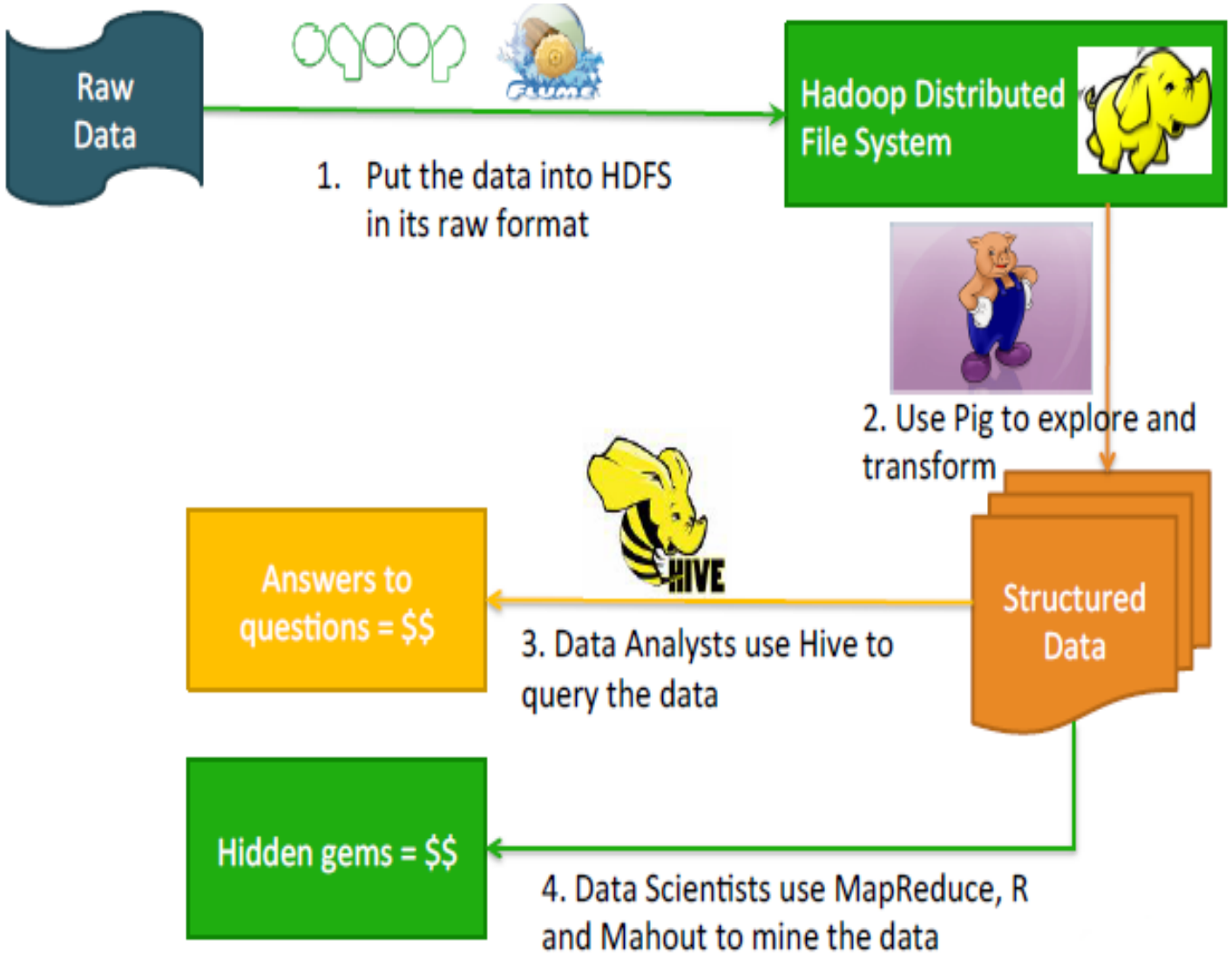






Hive advantages and limitations

- Advantages
  - Data warehouse - data analysis
  - Long running queries.
  - Fault tolerant environment. → on Hadoop
- Limitations
  - Slower response time. → due to MR
  - Data manipulation is not supported (fully).
- Applications
  - Batch processing (SQL based)
  - ETL jobs
  - Business Intelligence (Reports)
  - Predictive Modeling
  - Data mining
  - Log processing



Traditional ETL vs Hadoop ELT

- ETL stands for Extract, Transform and Load.
- The ETL process typically extracts data from the source/transactional systems, transforms it to fit the model of data-warehouse and finally loads it to the data warehouse.
- The transformation process involves cleansing, enriching and applying transformations to create desired output.
- Data is usually dumped to a staging area after extraction.

- ELT stands for Extract, Load and Transform.
- As opposed to loading just the transformed data in the target systems, the ELT process loads the entire data into the data lake. This results in faster load times.
- The load process can also perform some basic validations and data cleansing rules.
- The data is then transformed for analytical reporting as per demand.

Hive Installation & Getting started

- Install Hadoop.
- Install Hive
  - hive-site.xml
  - set PATH in ~/.bashrc
- Start metastore service.
- Start hive CLI.
- Start hiveserver2 service.
- Start hive beeline.

Hive QL

- Hive QL is extended SQL.
- Supports DQL, DML, DDL and DCL.
- DQL supports filtering, ordering, grouping, joins, etc. SELECT
  - Data will be read from HDFS.
  - Store query result into HDFS (or in another table).
- Supports views and indexes → deprecated in hive3.x → added materialized view in hive3.x
- Manage tables, partitions & buckets
- Provide various hive data types → primitive, collection (array, map, struct)
- Follows Schema on Read for better performance
  - While loading the data no schema is verified. (LOAD DATA ...)
  - While processing individual records schema is verified. (SELECT, INSERT, UPDATE, ...)
  - If data is not compatible with the type, value is considered null.

Hive data types

- Primitive Types:
  - BOOLEAN (1)
  - Integers: TINYINT (1), SMALLINT(2), INT(4), BIGINT(8)
  - Floating Point: FLOAT (single precision), DOUBLE (double precision), DECIMAL(m,n)
  - Characters: CHAR(n), VARCHAR(n), STRING
  - Date & Time: TIMESTAMP, DATE, DATETIME
- Collection Types:
  - ARRAY: collection of same type of data
  - STRUCT: collection of different type of data
  - MAP: collection of key-value pairs

Hive INSERT

- Inserts new records into hive table.
- Internally creates new files under HDFS (table directory).
- Produce MR job to insert data.
- While INSERT hive follows schema on write.

Hive SerDe

- Serde is Serializer & Deserializer.
  - Internally encapusulate Hadoop InputFormat (& RecordReader) and OutputFormat (& RecordWriter).
  - Types: Built-in Serdes (e.g. OpenCSVSerde), Third party Serdes, Custom Serdes

- OpenCSVSerde
  - Loads CSV file into hive table
    - Comma separated file
    - If data contains comma, cell is enclosed in double quote.
    - If data contains double quotes, it is escaped by "\".

- RegexSerde
  - Only Deserializer i.e. only used to read records.
  - Mainly used for data cleansing/extraction.

Thank you!

Nilesh Ghule <nilesh@sunbeaminfo.com>