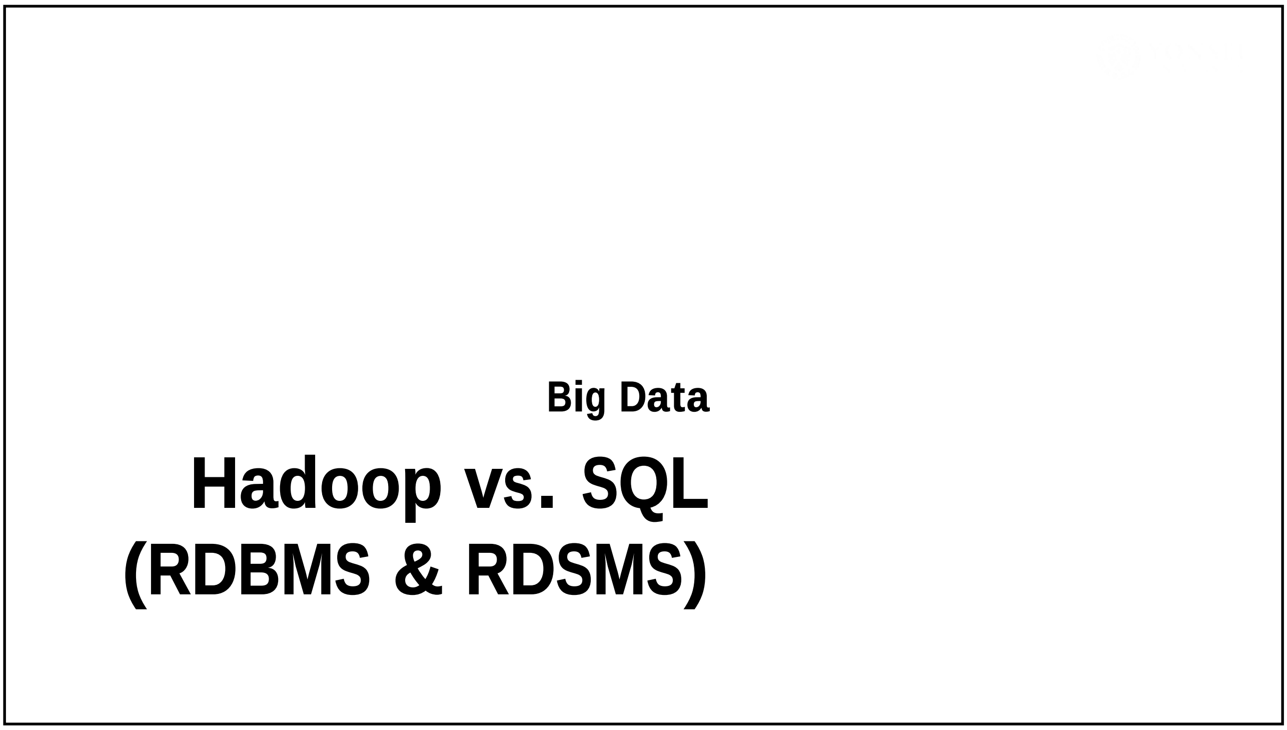
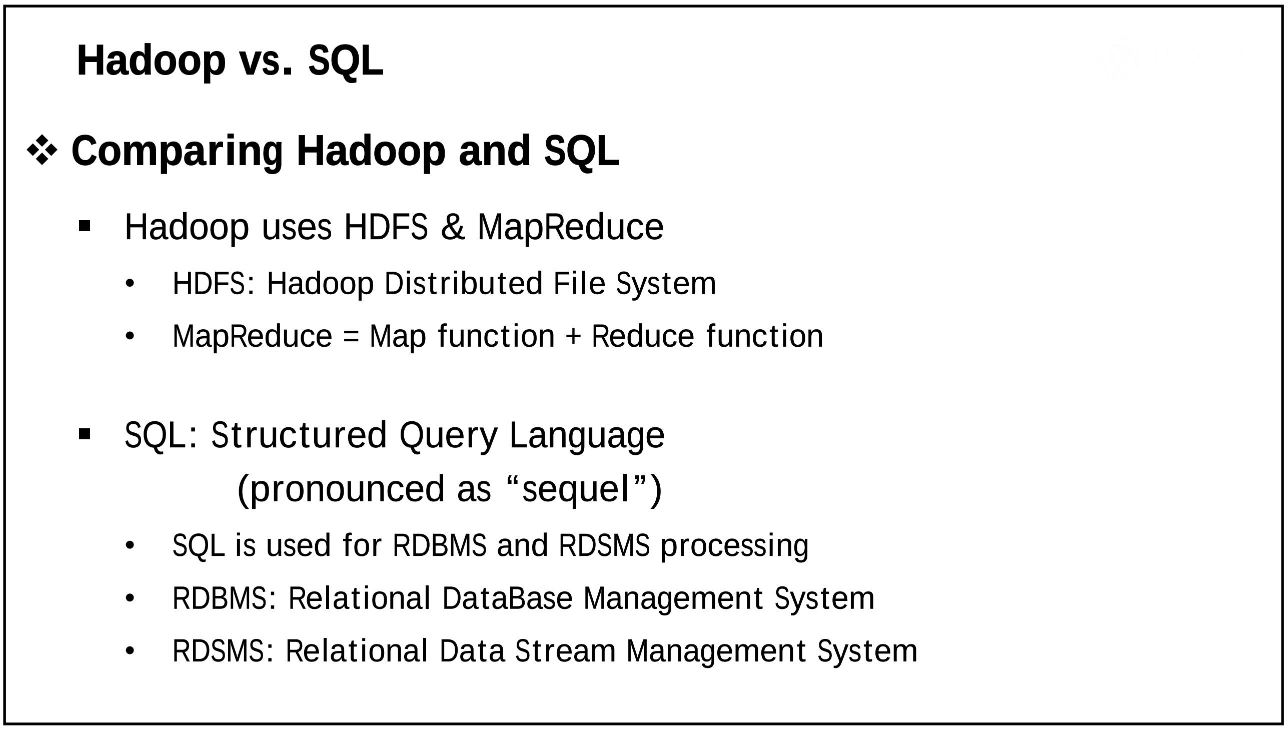


Big Data

Hadoop vs. SQL (RDBMS & RDSMS)



Hadoop vs. SQL

❖ Comparing Hadoop and SQL

- Hadoop uses HDFS & MapReduce
  - HDFS: Hadoop Distributed File System
  - MapReduce = Map function + Reduce function
- SQL: Structured Query Language
  (pronounced as "sequel")
  - SQL is used for RDBMS and RDSMS processing
  - RDBMS: Relational DataBase Management System
  - RDSMS: Relational Data Stream Management System

Hadoop vs. SQL

❖ Comparing Hadoop and SQL

- Schema on Read (Hadoop)
  vs.
  Schema on Write (SQL)
  - Schema = Schematic = Representation
    = Outline = Diagram

Hadoop vs. SQL

❖ SQL: Schema on Write

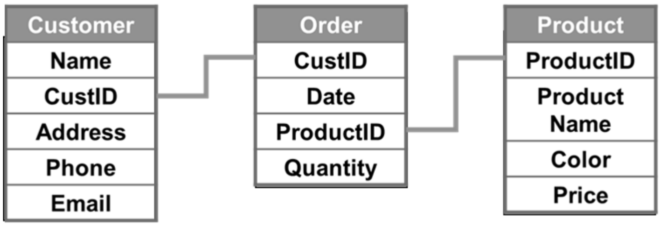
- Data structure must be known in advance and properly formatted for the DB write in process (i.e., recording, transfer, or combining)
- All analysis processing DB parts need to be fully completed, and then the distributed processed parts can be collected and combined

Hadoop vs. SQL

❖ SQL: Schema on Write

- Data is stored in a logically organized format (e.g., database metadata structure)

- Example: Writing data into an Excel file



Customer
Name
CustID
Address
Phone
Email

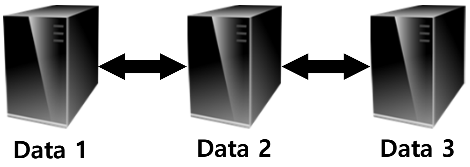
Order
CustID
Date
ProductID
Quantity

Product
ProductID
Product Name
Color
Price

Hadoop vs. SQL

❖ SQL: Schema on Write

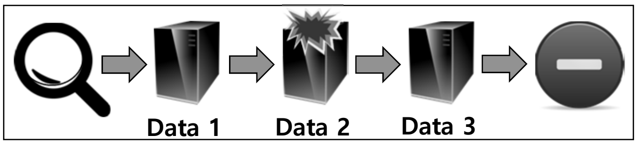
- Two Phase Commit
  - Distributed algorithm that makes a decision on a Transaction to Commit or Abort (Roll Back)
    - Example: Credit card transaction



Hadoop vs. SQL

❖ SQL: Schema on Write

- If one server node is delayed, the entire data analysis report will be delayed



Hadoop vs. SQL

❖ SQL: Schema on Write

- SQL is good in obtaining accurate results based on analysis of a DB of completed transactions
  - Example: Bank transactions DB analysis after the bank closes in the afternoon

Hadoop vs. SQL

❖ Hadoop: Schema on Read
= WORM (Write Once Read Many)

- Data inserted into HDFS does not need any preformatting and can have any structure
  - HDFS supports Unstructured data write in
- When reading the data from HDFS, the rule and structure is applied to the reading program code (e.g., Java program) that reads and analyzes the data

Hadoop vs. SQL

❖ Hadoop: Schema on Read
= WORM (Write Once Read Many)

- WORM ➔ Data is written in once to the HDFS, but it is read out multiple times using different search and MapReduce programs
- Data is divided and duplicated and then stored in the HDFS servers
  - Hadoop's default replication factor is 3

Hadoop vs. SQL

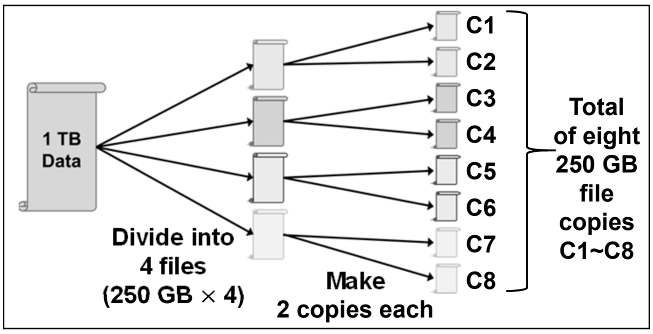
❖ **Hadoop: Schema on Read**
= WORM (Write Once Read Many)

- NameNode records where (on which HDFS DataNodes) the duplicated files are stored and processed with MapReduce

Hadoop vs. SQL

❖ **Hadoop: Schema on Read Example**

1. A 1 TB web sites search file is divided into 4 files and each 250 GB file is duplicated twice

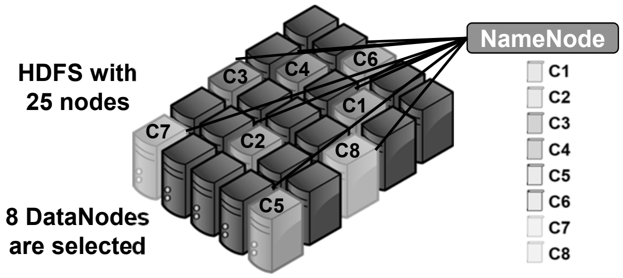

1 TB Data
Divide into 4 files (250 GB × 4)
Make 2 copies each
C1
C2
C3
C4
C5
C6
C7
C8
Total of eight 250 GB file copies C1~C8

▶ Hadoop's default replication factor is 3

Hadoop vs. SQL

❖ Hadoop: Schema on Read Example

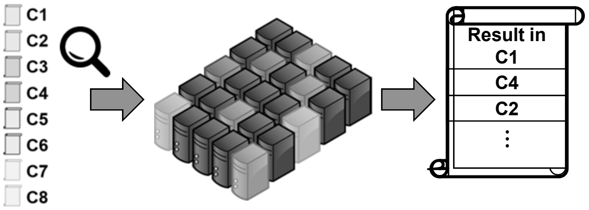
2. HDFS 25 node cluster will use 8 DataNodes to process these 8 files C1~C8 (one on each node)
3. NameNode records the 8 assigned DataNodes



Hadoop vs. SQL

❖ Hadoop: Schema on Read Example

4. A keyword search Java program is applied to all 8 DataNodes and a set of keywords are searched simultaneously by all 8 DataNodes
   - MapReduce is applied to different parts of the 1 TB file, so each DataNode server will need to analyze a 250 GB portion of the 1 TB file



Hadoop vs. SQL

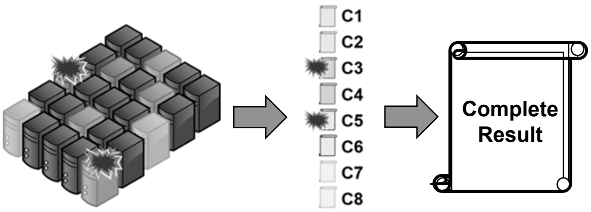
❖ Hadoop: Schema on Read Example

5. Map program searched results are delivered to the Reducer program (on the DataNodes)
   - MapReduce transforms the analysis problem into a computation process that uses a set of keys and values
     - <Key1, Value1>, <Key2, Value2>,…
6. For data analysis results, all distributed computing parts do NOT need to be fully completed to be combined

Hadoop vs. SQL

❖ Hadoop: Schema on Read Example

7. If a DataNode breaks down or is delayed, the data analysis results of the other completed parts will be analyzed and reported first, and later when the delayed parts get done, an update report will be made
   - Opposite to SQL's Two Phase Commit method



Hadoop vs. SQL

❖ Hadoop: Schema on Read Example

8. Good for databases that are continuously collecting new information (in various data types) and have to be consistently updated and analyzed

   Examples
   - Continuous collecting of new data
     - Websites & Emails
     - Social Networks
     - AR (Augmented Reality) systems
     - Keyword searches, etc.

Big Data

References

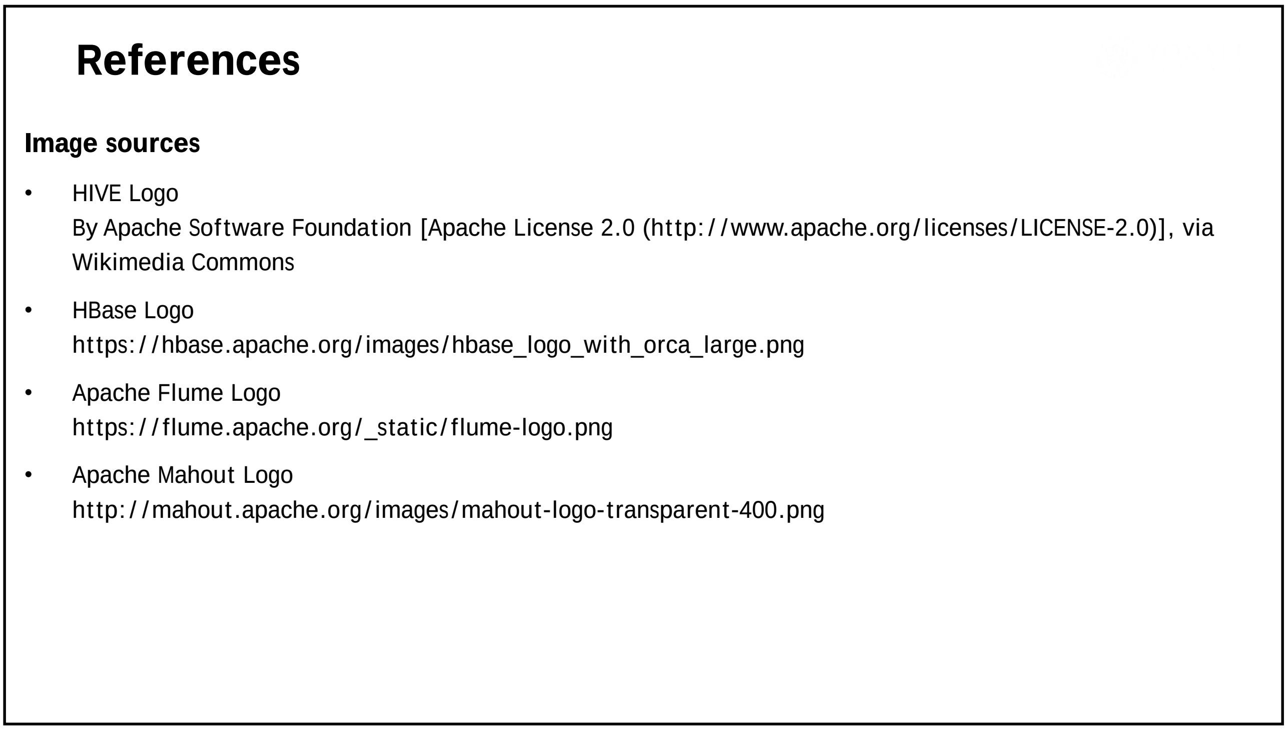
References

- I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. book in preparation, MIT Press, www.deeplearningbook.org, 2016.
- D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel & S. Dieleman, "Mastering the game of Go with deep neural networks and tree search," Nature, vol. 529, no. 7587, pp. 484-489, 28 Jan. 2016.
- N. Buduma, Fundamentals of Deep Learning: Designing Next-Generation Machine Intelligence Algorithms, O'Reilly Media, Jun. 2015.
- J. Heaton, Artificial Intelligence for Humans, Volume 3: Deep Learning and Neural Networks, Heaton Research, Inc., Nov. 2015.
- Jared Hillam, "What is Hadoop?: SQL Comparison," YouTube, https://www.youtube.com/watch?v=MfF750YVDxM
- Wikipedia, http://www.wikipedia.org

References

Image sources

- ORACLE Logo
  By Oracle Corporation. Cristan at en. wikipedia [Public domain],
  from Wikimedia Commons
- SAP Logo
  By SAP AG [Public domain], via Wikimedia Commons
- Microsoft Dynamics Logo
  http://news.microsoft.com/wp-content/uploads/2013/07/DynamicsLogoVertical_Web.jpg
- Hadoop Logo
  By Apache Software Foundation [Apache License 2.0 (http://www.apache.org/licenses/LICENSE-2.0)], via Wikimedia Commons



References

Image sources

- HIVE Logo
  By Apache Software Foundation [Apache License 2.0 (http://www.apache.org/licenses/LICENSE-2.0)], via Wikimedia Commons
- HBase Logo
  https://hbase.apache.org/images/hbase_logo_with_orca_large.png
- Apache Flume Logo
  https://flume.apache.org/_static/flume-logo.png
- Apache Mahout Logo
  http://mahout.apache.org/images/mahout-logo-transparent-400.png