Highly Scalable Databases

Lecture By
Binu Jasim
01-Nov-2016

BIG DATA

- E.g. website logs, phone data (what helps in many investigations), Large hadron collider experiment, Whatsapp/facebook data, Walmart/Amazon/ Flipkart, stock exchange data etc.
- Huge - orders of Tera bytes/Peta bytes/Exa bytes
- 3 V's of big data - Volume, Velocity, Variety

BIG DATA

- No definition of how big big data is
- Generally - Data that is difficult to be processed on a single machine
- Challenge 1 - How to store big data
- Challenge 2 - Analyse the huge (volume), fast streaming(velocity), often Schema less (variety) data

MapReduce

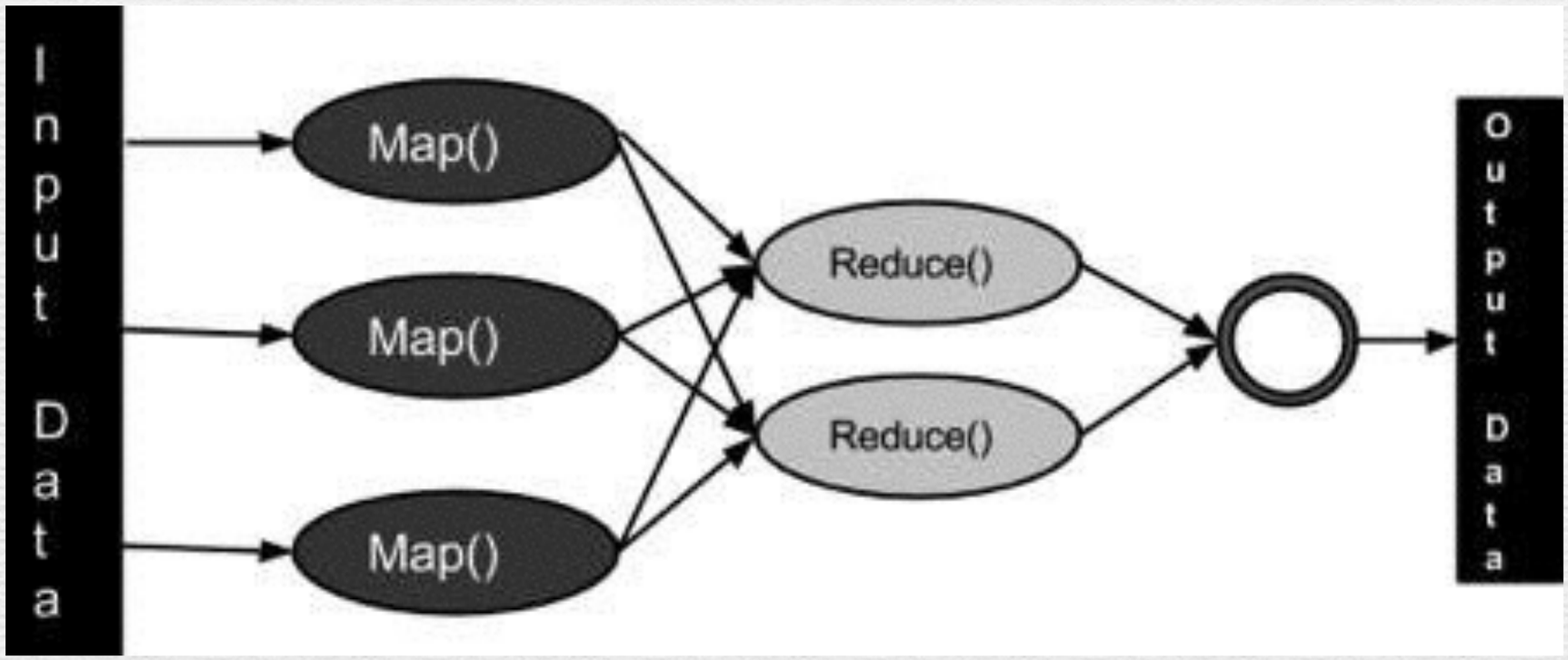
- Programming model for processing big data on clusters
- Invented by Google (Jeff Dean and Sanjay Ghemawat)
- Distributed parallel processing of data
- Used by Google to re-create their search indexes

- Jeff Dean jokes - (not for exam!)
- Jeff Dean writes directly in binary. He then writes the source code as documentation for other developers.
- Jeff Dean's PIN is the last 4 digits of pi
- Jeff Dean proved that P=NP. P=0 or N=1
- Jeff Dean took the bite out of Apple's logo.

Map Reduce Example: Find the total sales in each category across all stores of Walmart

Item	Cateogry	Price
Cabbage	Vegetables	35
Shampoo	Stationery	5
Oil	Grocery	82
Jacket	Textiles	652
Tomato	Vegetables	40

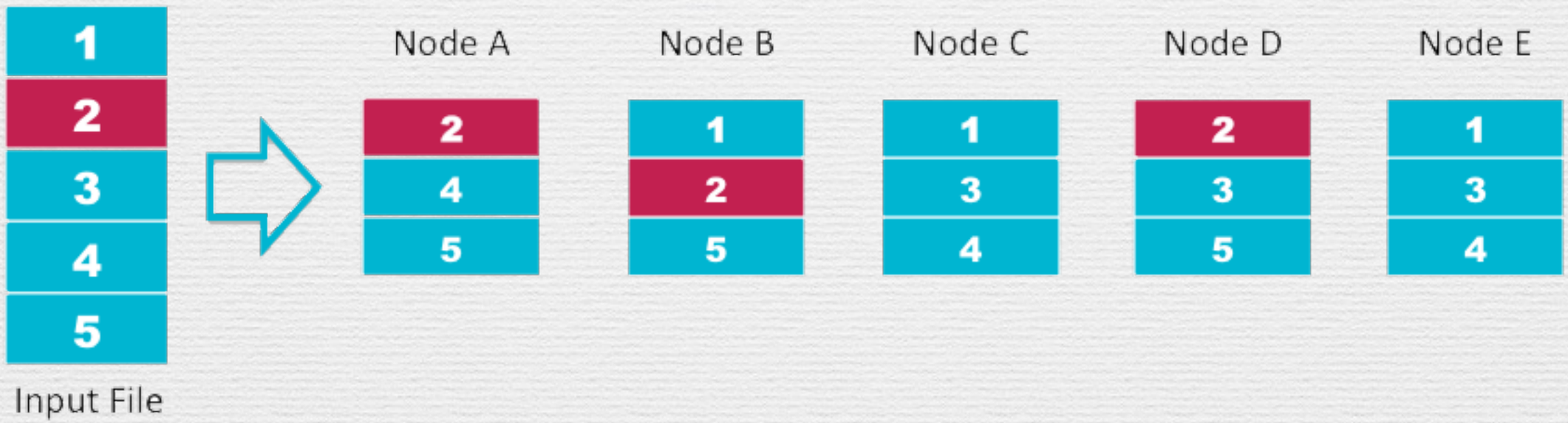
MapReduce



It is a very simple concept. The hard part is to manage the clusters, communicate between clusters, ensure fault tolerance etc.

- Open source implementation of MapReduce
- Developed By Dough Cutting (Yahoo)
- Uses Hadoop Distributed File System (HDFS)
- There are higher level abstractions over Hadoop such as Apache PIG, Apache Hive etc. (SQL based querying language)

HDFS



- Distributed Fault tolerant storage
- HDFS is a file system and not a DBMS

MapReduce

- Top N problem using MapReduce ?
- Each Mapper finds Top N and sends to a reducer
- Reducer finds the Top N out of all Top N's
- How to do word count using MapReduce?

MapReduce

- operates exclusively on <key, value> pairs
- The framework views the input to the job as a set of <key, value> pairs
- Produces a set of <key, value> pairs as the output of the job
- Not every algorithms can be implemented in MapReduce way

NoSQL Databases

NoSQL

- NoSQL - Not Only SQL (Not using the relational model)
- Old days: Hierarchical databases (1960s), Object databases (1990s) etc. But NoSQL used for -> Databases
- Running on clusters
- Mostly open-source
- Built for the 21st century web applications
- Schema-less

Document Database	Graph Databases
   	 
Key Value Stores	**Column Oriented Databases**
   	    

kinds of NoSQL databases

- **Key-value stores**: Redis, Riak, Voldemort
- **Column stores**: Cassandra, HBase
- **Document stores**: MongoDB, CouchDB, MarkLogic
- **Graph databases**: Neo4J

NoSQL

- Allied Market Research predicts that the global NoSQL market will reach $4.2 billion by the end of 2020

- The reason for this growth is that companies are looking for database technologies that offer greater flexibility, scalability and customisation for their applications.

- Eg. Guardian uses NoSQL to store news articles

Scale Horizontally

- Distribute across multiple clusters (Scale out)
- Introduces network latency, affected by network failures
- Can’t ensure Consistency - but ensures eventual consistency

Denormalization

- Denormalization and duplication of data
- Avoid Joins - data could be in different machines
- Trade off between storage space and speed
- Example - Keep the profile information with the every post of a user

Denormalization

- Scores(SSN, sName, SAT, ACT)
- Multiple ACT and SAT scores. Not in 4NF
- Can normalize into 3 relations R1(SSN, sName)
- R2(SSN, SAT) & R3(SSN, ACT)
- But if our queries always ask for a combined score - better not to normalize

Scalability

- Major reason for NoSQL databases
- RDBMS - Scale Up - No network latency but limited scaling
- NoSQL - Scale Out - distributed storage
- NoSQL used by Facebook messenger (Cassandra and HBase), LinkedIn, Google, Amazon… Almost everyone
- Google Cloud Datastore, Amazon's DynamoDB
- LinkedIn Voldemort

Sharding vs Replication

- Sharding in Reddit - comments in one machine, links/posts in another machine etc.
- Replication - same data is stored in multiple machines
- NoSQL databases like Cassandra support both

Difference Hadoop and NoSQL?

- Hadoop is for offline application - data analysis
- NoSQL - online response

1. Key Value Stores

- Key Value pairs - So only support simple querying by keys
- The database doesn't care about the value - blobs
- Application: instead of cache
- E.g. Redis drives Timeline in Twitter.
- Timeline is an index of tweets indexed by an id. The User Timeline, which consists of tweets the user has tweeted.

2. Document databases

```
{
  ISBN: 9780992461225,
  title: "JavaScript: Novice to Ninja",
  author: "Darren Jones",
  format: "ebook",
  price: 29.00
}
```

- Collections - similar to Tables in RDBMS
- Documents - similar to Rows in RDBMS

Tesla Model S

Overview	
Manufacturer	Tesla Motors
Also called	Code name: WhiteStar[1][2][3]
Production	2012–present
Model years	2012–present
Assembly	United States: Fremont, California (Tesla Factory) Europe: Tilburg, The Netherlands (all parts)
Designer	Franz von Holzhausen
Body and chassis	
Class	Full-size luxury (USA) Sports car (EU)
Body style	5-door liftback
Layout	Rear-motor, rear-wheel drive Dual motor all-wheel drive (*D* models)
Related	Tesla Model X
Powertrain	
Electric motor	Front and rear motor combined output up to 762 bhp (568 kW), 687 ft·lb (931 N·m), 3-phase AC induction motor
Transmission	1-speed fixed gear (9.73:1)

Rajdoot 350

Manufacturer	Escorts Group
Production	1983 - 1989
Predecessor	Yamaha RD350
Successor	Yamaha RX 100
Class	Standard
Engine	347 cc (21.2 cu in) Two stroke, Air-cooled, Parallel twin, twin carburetor, 7 port torque induction with reed valves
Ignition type	CB points
Transmission	6-Speed
Suspension	Front: Telescopic fork, Rear: Swingarm
Brakes	180mm Drum brakes (TLS front)
Tires	Front: 3.00-18"(4 ply rating), Rear: 3.50-18"(4 ply rating)
Wheelbase	1320 mm
Dimensions	L: 2040 mm W: 835 mm H: 1110 mm
Weight	143 kg (dry) 155 kg (wet)
Fuel capacity	16 Litres
Related	Yamaha RD350LC

Hindustan Ambassador Classic

Overview	
Manufacturer	Hindustan Motors
Production	1958–2014
Assembly	Uttarpara, Kolkata, West Bengal, India
Body and chassis	
Class	Compact car
Body style	4-door saloon
Layout	FR layout
Related	Morris Oxford series III
Powertrain	
Transmission	5-speed manual
Chronology	
Predecessor	Hindustan Landmaster

Wikipedia Info boxes see: DBPedia

MongoDB

- Document oriented NoSQL database
- Uses JSON for storing data (documents)
- One of the most popular NoSQL databases
- Supports an SQL like querying language

JSON

- Java Script Object Notations
- Alternative to XML (more popular than xml)
- Native support in many programming languages like python, javascript etc. - correspond to dicts

JSON

```
{"menu": {
  "id": 123,
  "value": "File",
  "popup": {
    "menuitem": [
      {"value": "New", "onclick": "CreateNewDoc()"}
      ,
      {"value": "Open", "onclick": "OpenDoc()"},
      {"value": "Close", "onclick": "CloseDoc()"}
    ]
  }
}}
```

Basically a nested hash map

Data modeling in Document Stores

```
{
 "sessionId": "session1",
 "speakers" : [
    {
        "id": 1, "name": "Alice", "pic": "..."
    },
    {
        "id": 2, "name": "Bob", "pic": "..."
    }]
}
```

Embed or Reference?

```
{
 "sessionId": "session1",
 "speakers" : [{"id":1}, {"id":2}]
},
```

```
{
    "id": 1, "name": "Alice", "pic": "..."
},
{
    "id": 2, "name": "Bob", "pic": "..."
}
```

What does MongoDB not being ACID compliant really mean?

- Atomic modifiers in MongoDB can only work against a single document.
- If you need to remove an item from inventory and add it to someone's order at the same time - you cant. Unless those two things - inventory and orders - exist in the same document
- So for transaction critical applications like Banks, RDBMS is the better choice.

3. Column oriented Databases

- Also called columnar databases/Column store
- Stores data tables as columns rather than as rows
- Inverse of a standard databases (RDBMS)
- Very high performance and a highly scalable
- Examples include: HBase, BigTable, Cassandra

RowId	EmpId	Lastname	Firstname	Salary
1	10	Smith	Joe	40000
2	12	Jones	Mary	50000
3	11	Johnson	Cathy	44000
4	22	Jones	Bob	55000

10:001,12:002,11:003,22:004;
Smith:001,Jones:002,Johnson:003,Jones:004;
Joe:001,Mary:002,Cathy:003,Bob:004;
40000:001,50000:002,44000:003,55000:004;

Apache Cassandra

- Cassandra is a distributed database management system
- It was developed at Facebook for inbox search.
- It was made an Apache top-level project since February 2010.
- It is a column-oriented database
- Used at Facebook, Twitter, Cisco, Rackspace, ebay, Twitter, Netflix, and more

ACID transactions

- Atomicity - Atomicity requires that each transaction be "all or nothing”
- Consistency - The consistency property ensures that any transaction will bring the database from one valid state to another.
- Isolation - No interference between transactions
- Durability - after transaction, effect persists

CAP Theorem

- CAP - Consistency, Availability, Partition tolerance
- Partition tolerance (the system continues to operate despite arbitrary partitioning due to network failures)
- Consistency (every read receives the most recent write or an error) - consistent across clusters
- Availability (every request receives a response, without guarantee that it contains the most recent version of the information)

CAP Theorem

- It is impossible to have all three at a time
- If distributed - either availability or consistency - not both
- In relational databases - can have both C & A, but it is not P

Visual Guide to NoSQL Systems

Availability:
Each client can always read and write.

A

Data Models

- Relational (comparison)
- Key-Value
- Column-Oriented/Tabular
- Document-Oriented

CA

RDBMSs (MySQL, Postgres, etc)

Aster Data
Greenplum
Vertica

AP

Dynamo
Voldemort
Tokyo Cabinet
KAI

Cassandra
SimpleDB
CouchDB
Riak

Pick Two

C

P

Consistency:
All clients always have the same view of the data.

CP

BigTable
Hypertable
Hbase

MongoDB
Terrastore
Scalaris

Berkeley DB
MemcacheDB
Redis

Partition Tolerance:
The system works well despite physical network partitions.

Exercise - Read Wikipedia articles about each of these

- Big Data
- Map Reduce, Hadoop, PIG, Hive
- NoSQL - key value stores, document databases, column stores, graph databases
- CAP theorem, Eventual consistency
- JSON, Document database data modeling

Reference

- Check cslab.org/dbms for practice questions in XPath and XQuery