Lectuer-5
Big Data Systems(SEZG522)

BITS Pilani
Pilani Campus

Slides: Courtesy:.Prof. Anindya

innovate achieve lead

BITS Pilani

Pilani Campus

First Semester

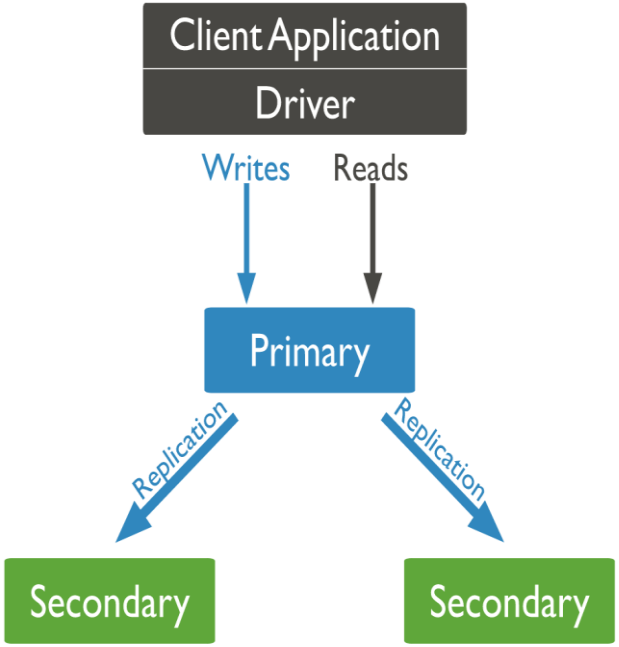
2022-23

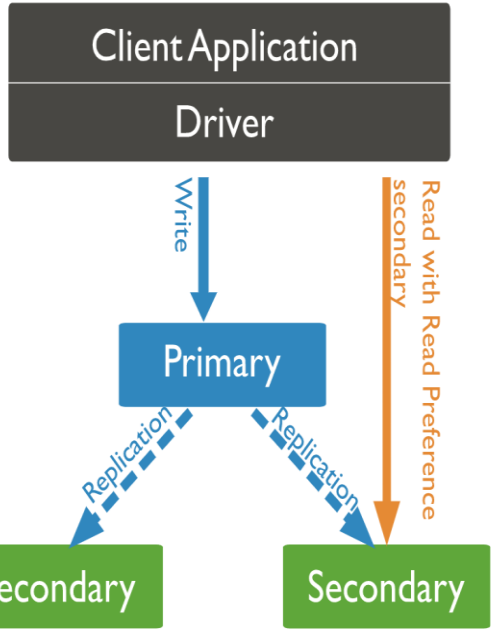
Lecture -5 Contents

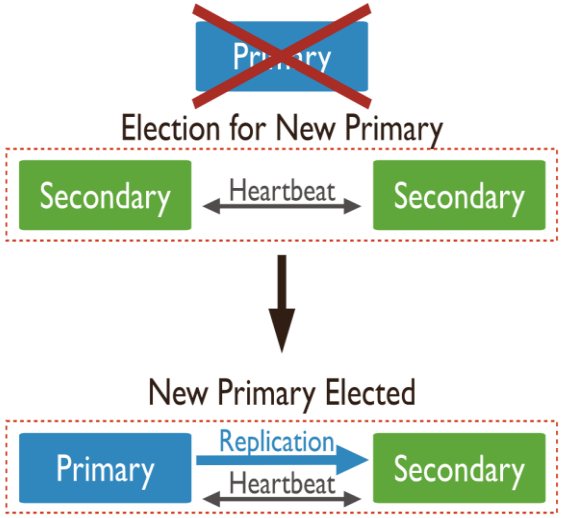
- Example BigData store options based on CAP requirement
  - ✓MongoDB
  - ✓Cassandra
- Big Data lifecycle

MongoDB

- Document oriented DB
- Various read and write choices for flexible consistency tradeoff with scale / performance and durability
- Automatic primary re-election on primary failure and/or network partition

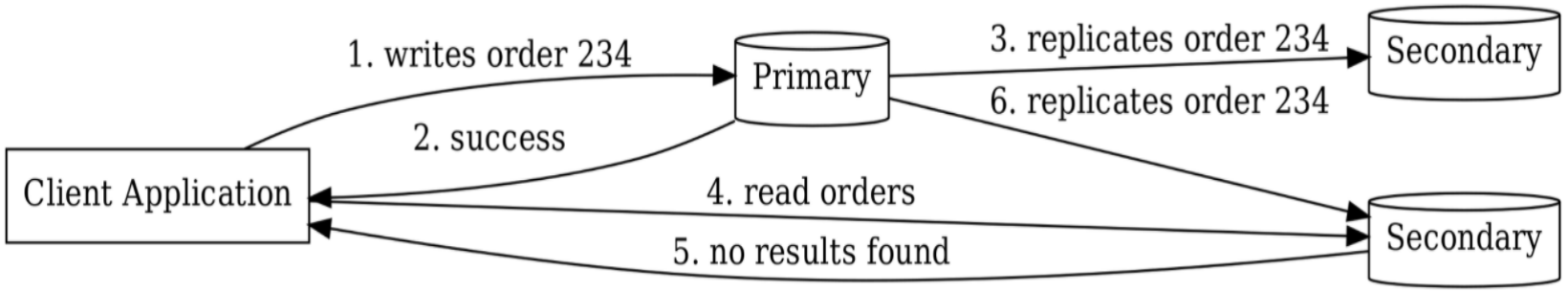




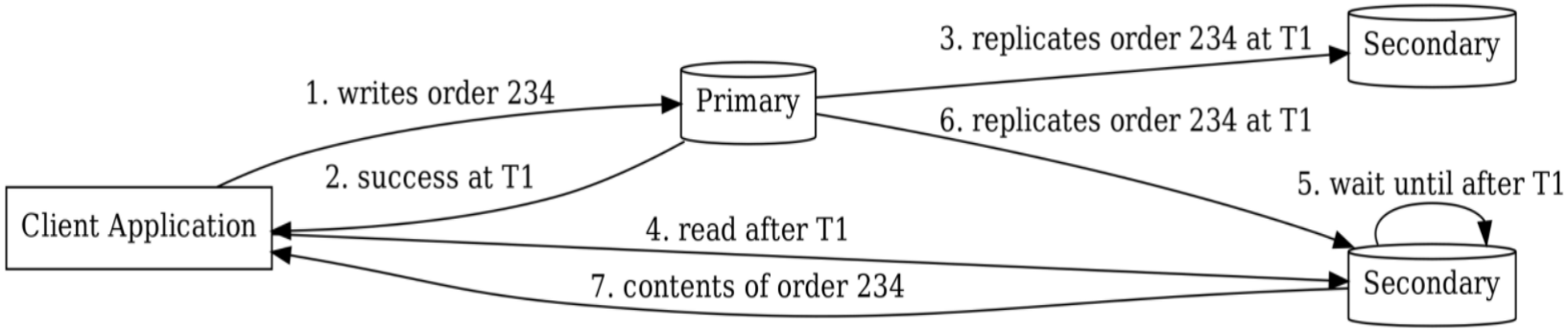


Example in MongoDB

Case 1 : No causal consistency



- Case 2: Causal consistency by making read to secondary wait



MongoDB “read concerns”

local :

- Client reads primary replica
- Client reads from secondary in causally consistent sessions

available:

- Read on secondary but causal consistency not required

majority :

- If client wants to read what majority of nodes have. Best option for fault tolerance and durability.

MongoDB “write concerns”

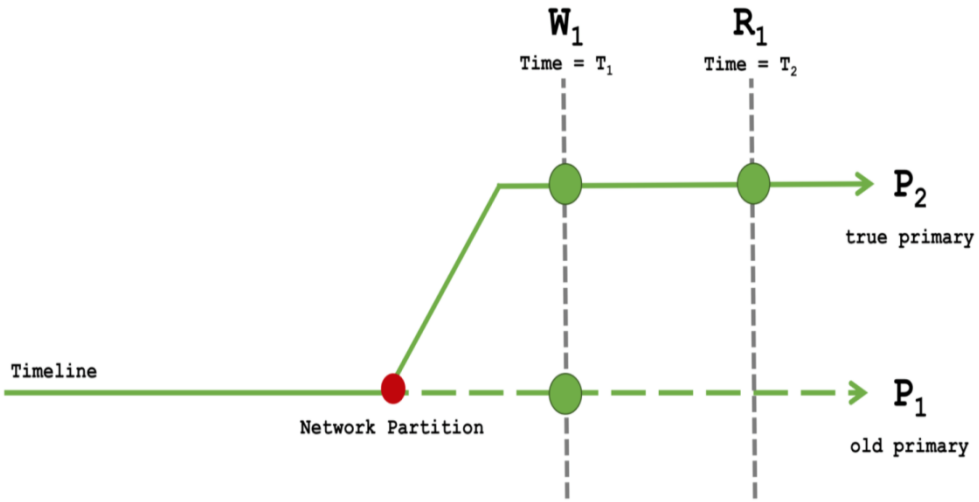
how many replicas should ack

- 1 - primary only
- 0 - none
- n - how many including primary
- majority - a majority of nodes (preferred for durability)

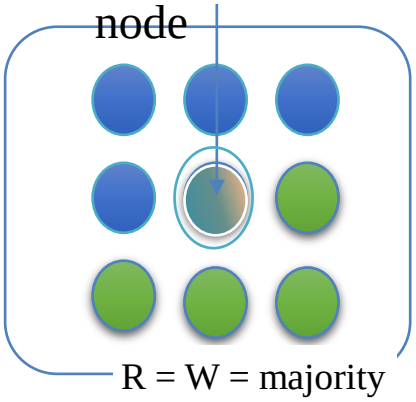
journaling - If True then nodes need to write to disk journal before ack else ack after writing to memory (less durable)

timeout for write operation

Consistency scenarios - causally consistent and durable



Read latest written value from common node



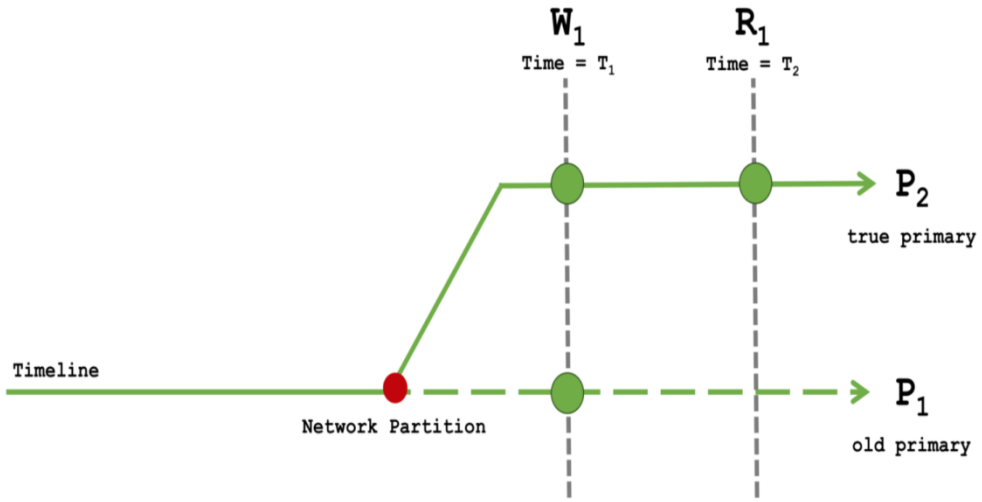
read=majority, write=majority

W1 and R1 for P1 will fail and will succeed in P2

So causally consistent, durable even with network partition sacrificing performance

Example: Used in critical transaction oriented applications, e.g. stock trading

Consistency scenarios - causally consistent but not durable



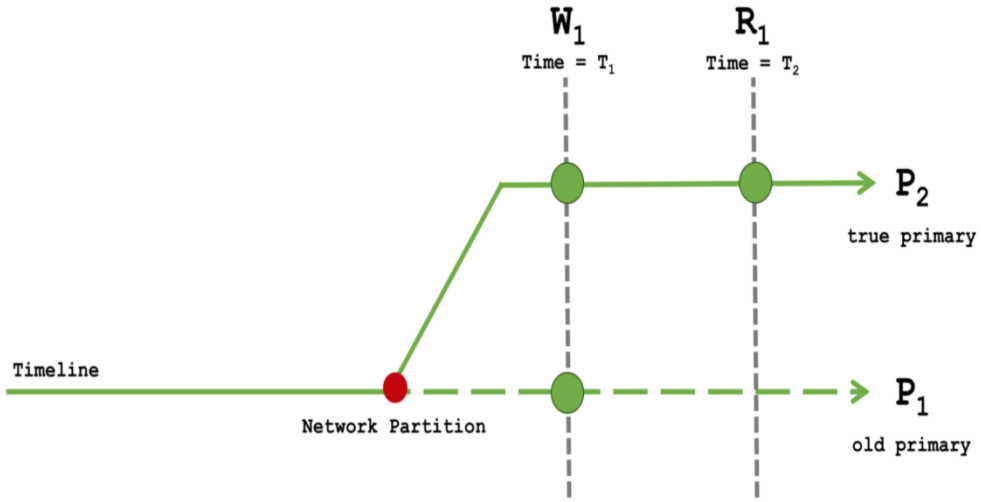
read=majority, write=1

W1 may succeed on P1 and P2. R1 will succeed only on P2. W1 on P1 may roll back.

So causally consistent but not durable with network partition. Fast writes, slower reads.

Example: Twitter - a post may disappear but if on refresh you see it then it should be durable, else repost.

Consistency scenarios - eventual consistency with durable writes

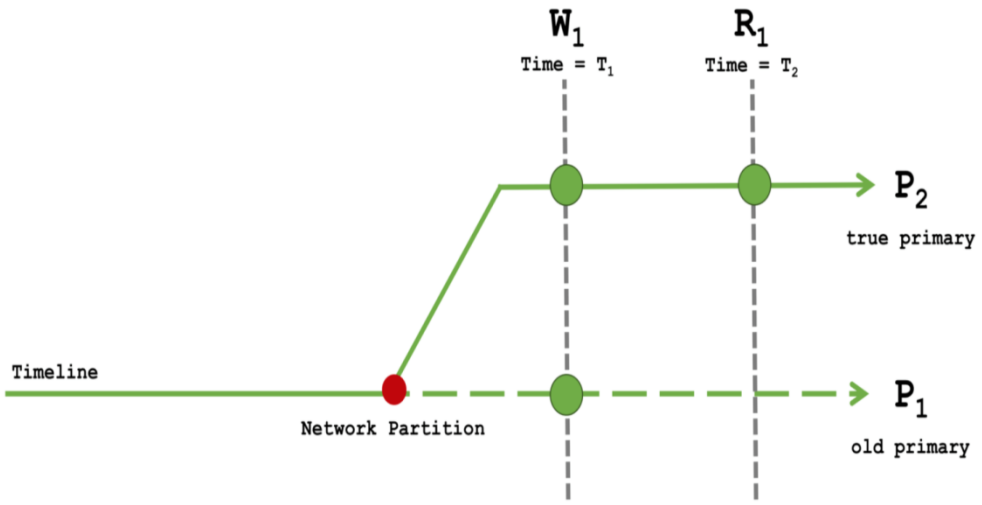


read=local, write=majority

W1 will succeed only for P1 and reads may not succeed to see the last write. Slow durable writes and fast non-causal reads.

Example: Review site where write should be durable but reads don't need causal guarantee as long as it appears some time (eventual consistency).

Consistency scenarios - eventual consistency but no durability



read=local, write=1

Same as previous scenario and not writes are also not durable and may be rolled back.

Example: Real-time sensor data feed that needs fast writes to keep up with the rate and reads should get as much recent real-time data as possible. Data may be dropped on failures.

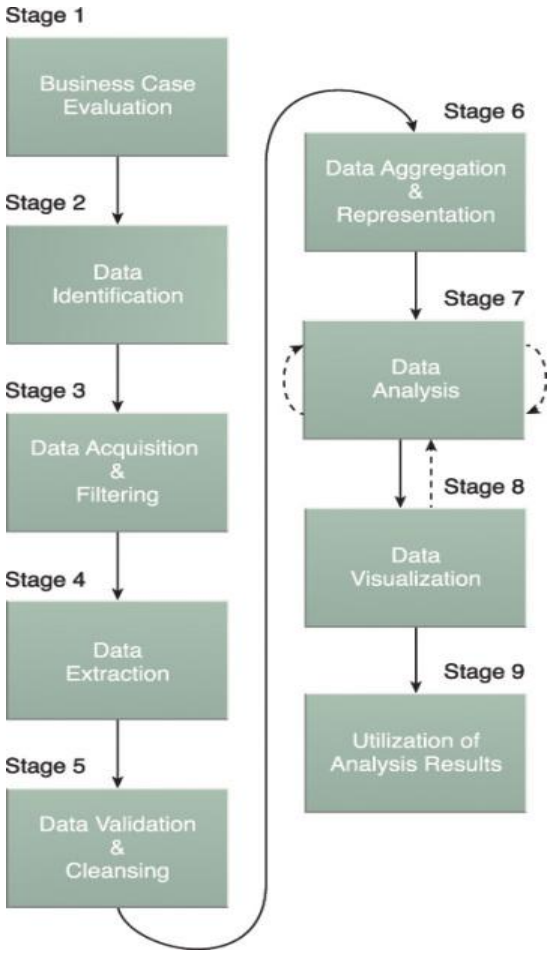
Big Data Analytics Lifecycle

- Big Data analysis differs from traditional data analysis primarily
  - ✓due to the volume, velocity and variety characteristics of the data being processes
- A step-by-step methodology is needed to organize the activities and tasks involved with
  - ✓Acquiring
  - ✓Processing
  - ✓Analyzing
  - ✓Repurposing data
- Explore a specific data analytics lifecycle that organizes and manages the tasks and activities associated with the analysis of Big Data

Big Data Analytics Lifecycle- Stages

Stages

1. Business Case Evaluation
2. Data Identification
3. Data Acquisition & Filtering
4. Data Extraction
5. Data Validation & Cleansing
6. Data Aggregation & Representation
7. Data Analysis
8. Data Visualization
9. Utilization of Analysis Results



1. Business Case Evaluation

Based on business requirements determine whether the business problems being addressed is really Big Data problem

- ✓ A business problem needs to be directly related to one or more of the Big Data characteristics of volume, velocity, or variety.

High volume
Unstructured data

Must begin with a well-defined business case that presents a clear understanding of the

- ✓ justification
- ✓ motivation
- ✓ goals of carrying out the analysis.

Find market fit for new product

A business case should be created, assessed and approved prior to proceeding with the actual hands-on analysis tasks.

What are the business questions ? Define thresholds on survey stats

Helps decision-makers to

- ✓ understand the business resources that will need to be utilized
- ✓ Identify which business challenges the analysis will tackle.
- ✓ Identify KPIs can help determine assessment criteria and guidance for the evaluation of the analytic results

2. Data Identification

Main objective is to identify the datasets required for the analysis project and their sources

- ✓Wider variety of data sources may increase the probability of finding hidden patterns and correlations.
  - ✓Caution: Too much data variety can also confuse - overfitting problem.
- ✓The required datasets and their sources can be internal and/or external to the enterprise.

Identify respondents
Demographics
What questions to ask
Do we need other surveys

For internal datasets

- ✓A list of available datasets from internal sources, such as data marts and operational systems, are typically compiled and verified for data required

For external datasets

- ✓A list of possible third-party data providers, such as data markets and publicly available datasets needs to be compiled
- ✓Data may be embedded within blogs or other types of content-based web sites, automated tools needs to be used to extract it

3. Data Acquisition & Filtering

The data is gathered from all of the data sources that were identified during the last stage

The acquired data is then looked upon for

- ✓ filtering / removal of corrupt data
- ✓ removal of unusable data for analysis

Clean bad data, e.g. empty responses
Junk text inputs
Filter a subset if we don't need to look at all attributes, all demographics

In many cases involving unstructured external data, some or most of the acquired data may be irrelevant (noise) and can be discarded as part of the filtering process.

"Corrupt" data can include records with missing or nonsensical values or invalid data types

- ✓ Advisable to store a verbatim copy of the original dataset before proceeding with the filtering

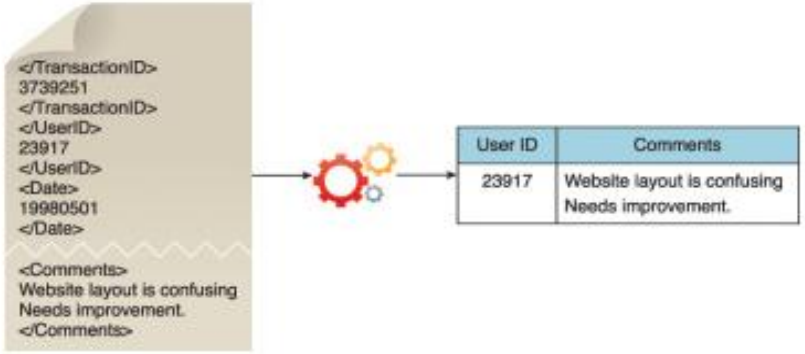
Data needs to be persisted once it gets generated or enters the enterprise boundary

- ✓ For batch analytics, this data is persisted to disk prior to analysis
- ✓ For real-time analytics, the data is analyzed first and then persisted to disk

4. Data Extraction

- Dedicated to extracting data and transforming it into a format that the underlying Big Data solution can use for the purpose of the data analysis
- The extent of extraction and transformation required depends on the types of analytics and capabilities of the Big Data solution.

Structure the unstructured response



Comments and user IDs are extracted from an XML document.

5. Data Validation & Cleansing

Validate survey responses
Contradictory answers
Identify population skews, e.g. responses have inherent gender bias so no point in making a gender based analysis
Codify certain columns for easier analysis

Invalid data can skew and falsify analysis results

Data input into Big Data analyses can be unstructured without any indicat validity

- ✓Complexity can further make it difficult to arrive at a set of suitable validation constraints
- ✓Dedicated stage is required to establish complex validation rules and removing any known invalid data.

Big Data solutions often receive redundant data across different datasets.

- ✓This can be exploited to explore interconnected datasets in order to
  - assemble validation parameters
  - fill in missing valid data

For batch analytics, data validation and cleansing can be achieved via an offline ETL operation

For real-time analytics, a more complex in-memory system is required to validate and cleanse the data as it arrives from the source

6. Data Aggregation & Representation

Final joined data set (e.g. current with old survey or 3rd party demographics data) with certain aggregations done for downstream analysis

Dedicated to integrating multiple datasets together to arrive at a unified view

- ✓ Needs to merge together the data spread across multiple datasets through common field
- ✓ Needs reconciliation of data coming from different sources
- ✓ Needs to identify the dataset representing the correct value needs to be determined.

Can be complicated because of :

- ✓ Data Structure – Although the data format may be the same, the data model may be different
- ✓ Semantics – A value that is labeled differently in two different datasets may mean the same thing, for example “surname” and “last name.”

The large volumes makes data aggregation a time and effort-intensive operation

- ✓ Reconciling these differences can require complex logic that is executed automatically without the need for human intervention

Future data analysis requirements need to be considered during this stage to help foster data reusability.

7. Data Analysis

Dedicated to carrying out the actual analysis task, which typically involves one or more types of analytics

- ✓ Can be iterative in nature, especially if the data analysis is exploratory
- ✓ Analysis is repeated until the appropriate pattern or correlation is uncover

Various types of descriptive / predictive analysis on survey data to understand market fit for new product. Writing SQL on data and create charts. Build models on the data for hypothesis testing, prediction.

Depending on the type of analytic result re

- ✓ Can be as simple as querying a dataset to compute an aggregation for comparison
- ✓ Can be as challenging as combining data mining and complex statistical analysis techniques to discover patterns and anomalies or to generate a statistical or mathematical model to depict relationships between variables.

Confirmatory / Exploratory data analysis

Confirmatory data analysis

- ✓A deductive approach where the cause of the phenomenon being investigated is proposed beforehand - a hypothesis
- ✓Data is then analyzed to prove or disprove the hypothesis and provide definitive answers to specific questions
- ✓Data sampling techniques are typically used
- ✓Unexpected findings or anomalies are usually ignored since a predetermined cause was assumed

Exploratory data analysis

- ✓Inductive approach that is closely associated with data mining
- ✓No hypothesis or predetermined assumptions are generated
- ✓Data is explored through analysis to develop an understanding of the cause of the phenomenon
- ✓May not provide definitive answers
- ✓Provides a general direction that can facilitate the discovery of patterns or anomalies