DATA ANALYTICS AND BIG DATA

Fundamentals of Artificial Intelligence

Session 16

Pramod Sharma
pramod.sharma@prasami.com

2

Agenda

- Introduction
- Data Analysis and Data Analytics
- Reference Architecture
- Big data platforms

pra-sâmi

3

Netflix

pra-sâmi

4

Who is Watching

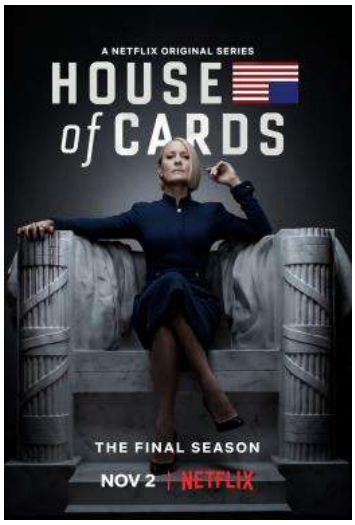
- In 2006 Netflix announced the Netflix Prize, a competition for creating an algorithm that would "substantially improve the accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences."
- Winning algorithm showed 10% improvement
- Netflix decided not to roll out in production after a thorough ROI

pra-sâmi

5

House of Cards

- Do you believe that the show ***House of Cards*** was chosen because Netflix "thought subscribers might like it." ?
- $100 million show wasn't green-lighted solely because it seemed like a good plot!
- Netflix is Data Driven Company,
  - All their decisions are based on data
- For US rights, they out-bid HBO, AMC and others
  - At a cost of $4 million to $6 million an episode, this resulted in 2-season price tag is over $100 million



pra-sâmi

6

House of Cards

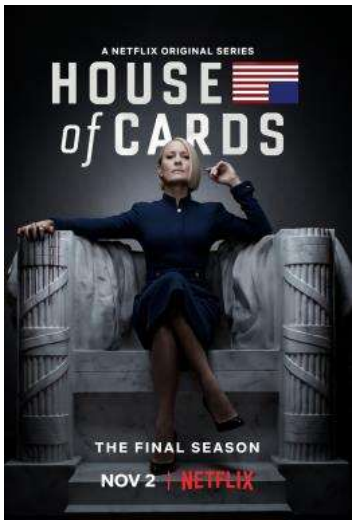
- Because Netflix knew:
  - A lot of users watched the David Fincher directed movie The Social Network from beginning to end
  - The British version of "House of Cards" has been well watched
  - Those who watched the British version "House of Cards" also watched Kevin Spacey films and/or films directed by David Fincher
- Having bought, and with the data they have,
  - Make a "personalized trailer" for each type of Netflix member, not a "one size fits all" trailer
  - Netflix made 10 different cuts of the trailer for House of Cards,
  - Each geared toward different audiences
- The trailer you saw was based on your previous viewing behavior
  - If you watched a lot of Kevin Spacey films, you saw a trailer featuring him
  - Those who watched a lot of movies starring females saw a trailer featuring the women in the show
  - David Fincher fans saw a trailer featuring his touch

pra-sâmi

7

House of Cards - Success Story

- It has brought in 2 million new U.S. subscribers in the first quarter of 2013
- A 7% increase over the previous quarter.
- It also brought in 1 million new subscribers from elsewhere in the world
- 3 million subscribers almost paid Netflix back for the cost of *House of Cards*.
- What about current subscribers?
  - Does having *House of Cards* make stick around?
- Yes, for 86% of them.



8

Data Analytics

Analytics Characteristics are not new

- Value: produced when the analytics output is put into action
- Veracity: measure of accuracy and timeliness
- Quality:
  - well-formed data
  - Missing values
  - cleanliness
- Latency: time between measurement and availability
- Data types have differing pre-analytics needs

9

The Real Time Boom..

Facebook Real Time Social Analytics

SaaS Real Time User Tracking

Google Real Time Web Analytics

Twitter paid tweet analytics

New Real Time Analytics Startups..

mixpanel

tctango

Google Real Time Search

pra-sâmi

10

Data Analysis and Data Analytics

pra-sâmi

11

Example of Analytics

- Researching
  - What features get re-tweeted
  - Duplicate detection
  - Sentiment analysis

Batch

- Correlating
  - Desktop vs Mobile user ?
  - What devices fail at the same time?
  - What features get user hooked?

Near Real Time

- Counting
  - How many request/day?
  - What's the average latency?
  - How many signups, sms, tweets?

Real Time

pra-sâmi

12

Skills Required for Big Data Analytics

- Store and process
  - Large scale databases
  - Software engineering
  - System/network engineering
- Analyze and model
  - Reasoning
  - Knowledge representation
  - Multimedia retrieval
  - Modeling and simulation
  - Machine learning
  - Information retrieval
- Understand and design
  - Decision theory
  - Visual analytics
  - Perception cognition

pra-sâmi

13

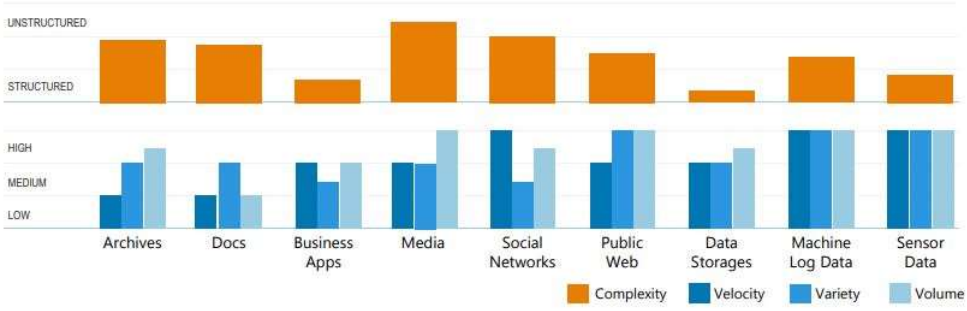
Relational vs. Non-Relational Architecture

Relational	Non- Relational
❑ Rational	❑ Agile
❑ Predictable	❑ Flexible
❑ Traditional	❑ Modern

pra-sâmi

14

Big Data Challenges



Archives
Scanned documents, statements, medical records, e-mails etc..

Media
Images, video, audio etc.

Data Storages
RDBMS, NoSQL, Hadoop, file systems etc.

Docs
XLS, PDF, CSV, HTML, JSON etc.

Social Networks
Twitter, Facebook, Google+, LinkedIn etc.

Machine Log Data
Application logs, event logs, server data, CDRs, clickstream data etc.

Business Apps
CRM, ERP systems, HR, project management etc.

Public Web
Wikipedia, news, weather, public finance etc

Sensor Data
Smart electric meters, medical devices, car sensors, road cameras etc.

pra-sâmi

15

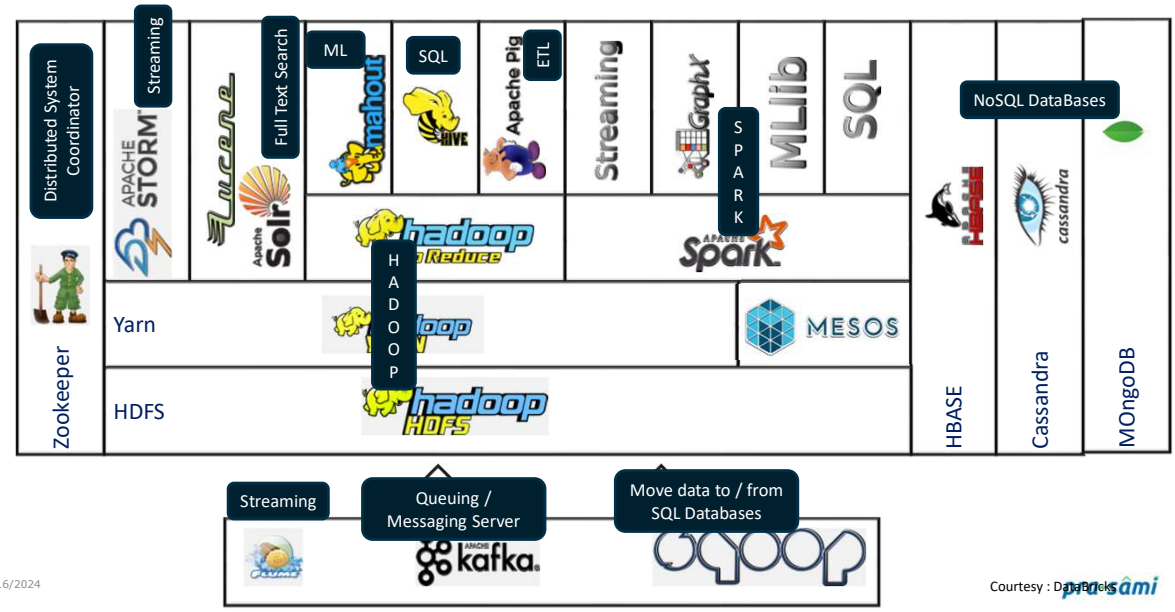
Big Data Analytics

	Traditional Analytics (BI)	Big Data Analytics
Focus on	✓ Descriptive analytics ✓ Diagnosis analytics	✓ Predictive analytics ✓ Data science
Data Sets	✓ Limited data sets ✓ Cleansed data ✓ Simple models	✓ Large scale data sets ✓ More types of data ✓ Raw data ✓ Complex data models
Supports	Causation: what happened, and why?	Correlation: new insight more accurate answers

pra-sâmi

17

Recall Hadoop Ecosystem



Courtesy : DataBricks

pra-sâmi

18

Hadoop MapReduce

- Hadoop MapReduce framework is primarily designed for batch processing
- Less suitable for:
  - Ad-hoc data exploration,
  - Machine learning processes
- Attempts are being made to replace MapReduce with alternatives
- In case of SQL on Hadoop, initiative that aims to improve Hive performance are:
  - Cloudera Impala,
  - Pivotal HAWQ
  - Hortonworks Stinger
- Only Java natively supported
  - Support for others languages needed

Apache Spark is another alternative to replace MapReduce with a more performing execution engine but still use Hadoop HDFS as storage engine for large data sets.

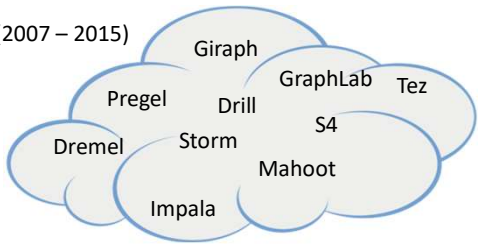
19

Hadoop

(2001 – 2013)

General Batch Processing

(2007 – 2015)



Specialized Systems
(Iterative, Interactive, ML, Streaming, Graph, SQL, etc.)

(2014 – ?)

General Unified Engine
One-stop shop

- No general purpose computing engine in the industry
  - Batch processing ➔Hadoop MapReduce
  - Perform stream processing ➔Apache Storm / S4
  - Interactive processing ➔Apache Impala / Apache Tez
  - Graph processing ➔Neo4j / Apache Giraph/ Pregel
- No powerful engine, that can process the data both in real-time and batch mode
- Need for sub-second response and perform in-memory processing

20

Birth of Spark

- 2009 – AmpLab – University of Berkley
  - This project is still alive
  - First published the following year in a paper entitled “Spark: Cluster Computing with Working Sets”
- AMP: Algorithms Machines People - Turning up the Volume on Big Data
  - Working at the intersection of three massive trends: powerful machine learning, cloud computing, and crowdsourcing, the AMPLab is creating a new Big Data analytics platform that combines Algorithms, Machines and People to make sense at scale
- Algorithms
  - Machine learning (ML) turns data into information and knowledge. While it is useful to view ML as a toolbox that can be deployed for many data-centric problems, our long-term goal is more ambitious—we are developing ML as a full-fledged engineering discipline
- Machines
  - Many claim that the “datacenter is the new computer” but datacenters do not provide the key services one needs for managing and understanding massive data. We are developing a scalable software platform to make using a datacenter for analytics as easy as using an individual computer today
- People : Leveraging Human Intelligence and Activity
  - People will play a key role in Big Data applications – not simply as passive consumers of results, but as active providers and gatherers of data, and to solve ML-hard problems that algorithms on their own cannot solve. The AMPLab is building tools that include people as individuals and crowds for all phases of the analytics lifecycle

21

Birth of Spark

- Researchers were working on Cluster Manager “Mesos”
- Mesos is similar to “Yarn”
  - Yarn is resource manager for Hadoop
  - It was introduced with Hadoop 2.0
- In those days there was no Yarn!
- In 2009, they created a testing harness to test the Mesos,
  - Everything was in-memory
  - Not only it tested Mesos, it also gave them idea to productize it
- In 2012, Spark was made open source
  - It became top Apache project
  - At the same time, founders also floated a company under name of “DataBricks”

MESOS

databricks

22

"Apache Spark is an open source big data processing framework built around speed, ease of use, and sophisticated analytics.

It was originally developed in 2009 in UC Berkeley's AMPLab, and open sourced in 2010 as an Apache project.

Its in-memory processing engine allowing efficient execution of streaming, ML or SQL workload."

pra-sâmi

version: 3.5.0

23

- Spark 0.9 2014-04-11
- Spark 1.0 2014-05-30
- Spark 1.6 2016-01-04
- Spark 2.0 2016-07-26
- Spark 2.1 2016-12-22
- Spark 2.4.8 2021-05-14
- Spark 3.0.0-preview/ 2019-11-06
- Spark 3.0.0-preview2/ 2019-12-22
- Spark 3.0 2020-06-16
- Spark 3.1.1 2021-03-02
- Spark 3.2.3 2022-11-28
- Spark 3.3.0 2022-06-16
- Spark 3.4.0 2023-04-13
- Spark 3.5.0 2023-09-13
- Spark 3.5.1 2024-02-23

Release1.6 was very **Stable**, Industry used it for some time even after new releases came.
The Scala side worked perfect.
Python side was slow though.

Release 2 onwards an abstraction layer was implemented, that made it fast on all supported languages

Spark 3 is about 2 times faster than release 2.x

Current Stable release

pra-sâmi

24

What is Spark?

- Can perform both batch processing and stream processing.
  - Previously collected job in a single batch
  - Deal with Spark streaming data as well
- Integrates with all the Big Data tools
  - Can access any Hadoop data source
  - Can run on Hadoop clusters
  - Extends Hadoop MapReduce to the next level, including iterative queries and stream processing.
- Spark is not an extension of Hadoop
  - Spark is independent of Hadoop,
  - Has its own cluster management system
  - Uses Hadoop for storage purpose only
- Spark is written in Scala
  - Offers rich APIs in Scala, Java, Python, as well as R

25

Features - Speed

- In-memory caching of data (for iterative, graph, and machine learning algorithms, etc.)
  - Runs computations in-memory
  - Up to 100 times faster while running in memory
  - Upto 10 times faster when running on disk
  - By reducing number of read/write operations
- Has an advance DAG execution engine that supports acyclic data flow and in-memory computing
- Has many other workflows, i.e. join, filter, flatMapdistinct, groupByKey, reduceByKey, sortByKey, collect, count, first…
  - around 30 efficient distributed operations
- Supports interactive shells for exploratory data analysis
- Spark API is extremely simple to use

26

Features - Generality

- A general programing model
  - Write an application by composing arbitrary operations
- Easy to combine different processing model seamlessly in the same example e.g.
  - Data classification through spark ML libraries
  - Streaming data through source via Spark Streaming
  - Querying the results in real time through spark SQL
- Supports multiple languages
  - Built-in APIs for Java, SQL, R, Scala, or Python
  - 80 high-level operators for interactive querying

27

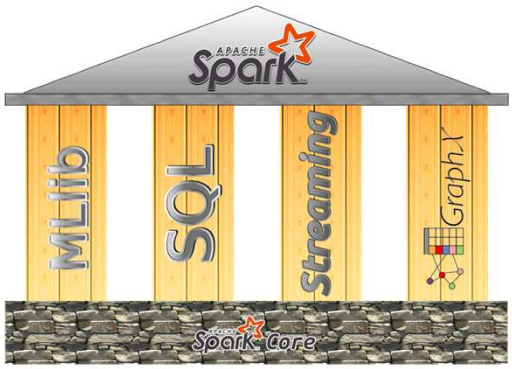
Features - Advance Analytics

- Spark not only supports :
  - 'Map' and
  - 'reduce'
- Also supports :
  - SQL queries
  - Streaming data
  - Machine learning (ML),
  - Graph algorithms

28

Apache Spark Components

- Spark Core
  - Execution platform for all the Spark applications
- Spark SQL
  - Can process structured as well as semi-structured data,
  - Runs unmodified queries up to 100 times faster on existing deployments
- Spark Streaming
  - Across live streaming enables an interactive and data analytics application
- Spark Mllib
  - Capable of in-memory data processing
- Spark GraphX
  - Graph computation engine

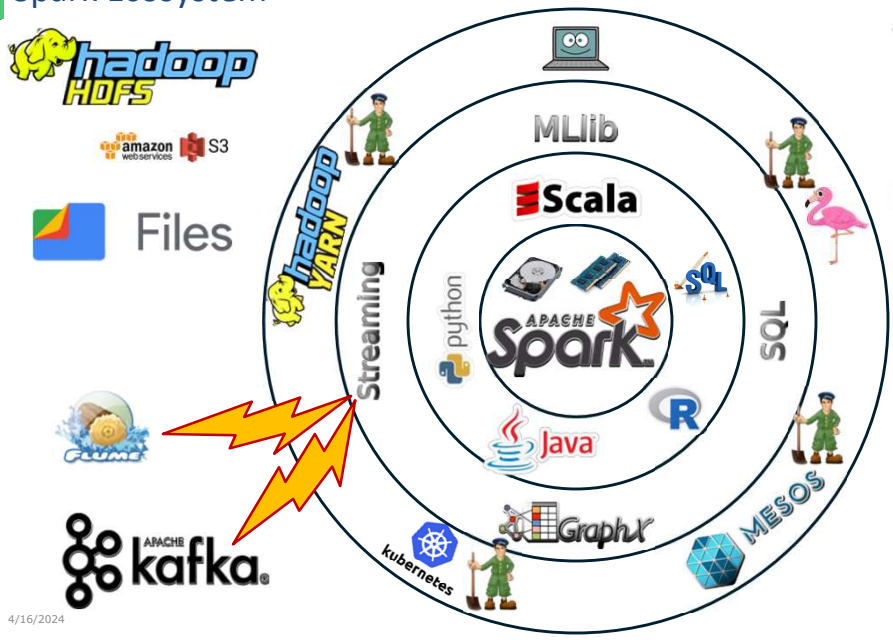


The four pillars of mana :
According to Polynesians, mana is the power of the elemental forces of nature embodied in an object or person.

pra-sâmi

29

Spark Ecosystem

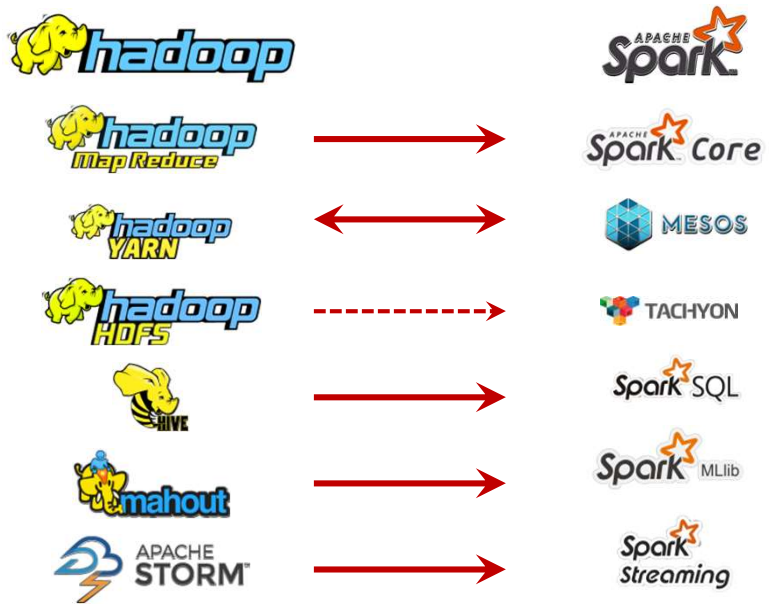


Courtesy : DataBricks

pra-sâmi

30

Survival of the Fittest

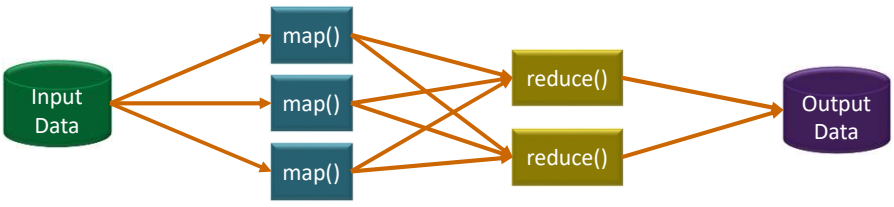


Courtesy : DataBricks

31

Hadoop MapReduce

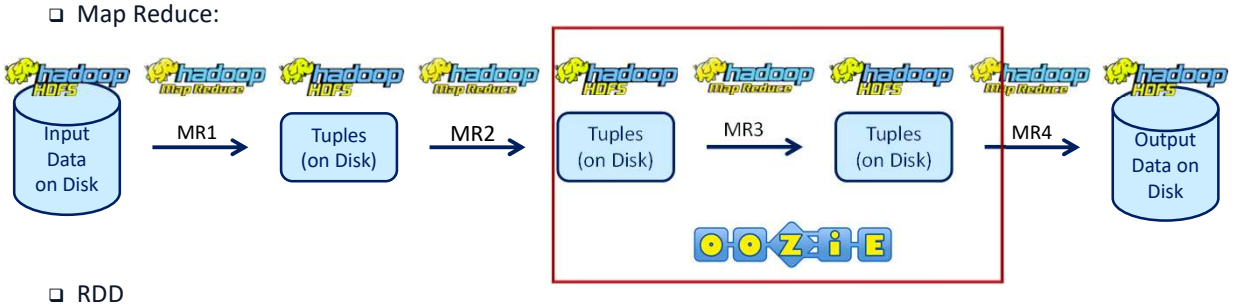
- The term MapReduce actually refers to two separate and distinct tasks that Hadoop programs perform. The first is the map job … and the second is the reduce.



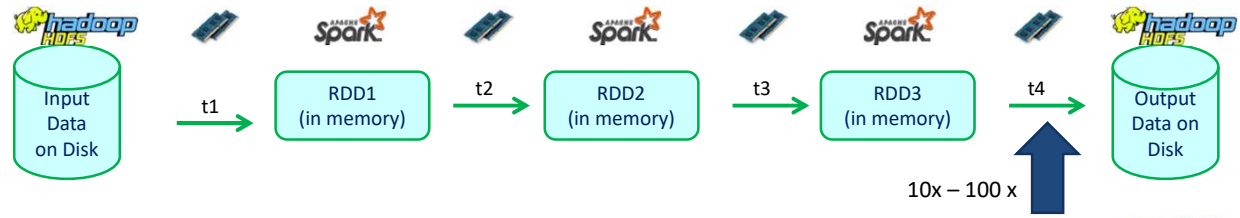
32

MapReduce and RDD Comparison

- Map Reduce:



- RDD



pra-sâmi

33

Big Data Architecture

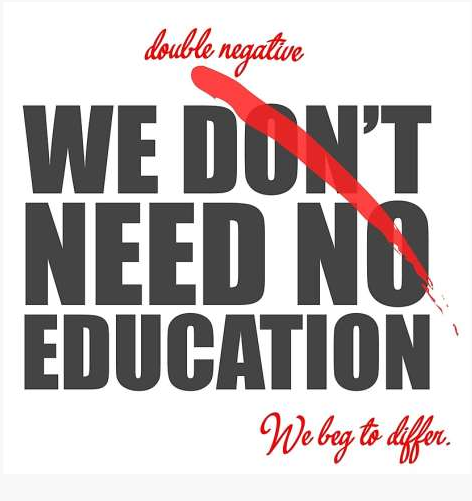
pra-sâmi

34 Need for New Technologies

- RDBMS are the square peg for all round holes
- Not all data requires ACID Compliant data stores
  - Atomicity, Consistency, Isolation, Durability
- To implement ACID, tradeoffs limit scalability of traditional systems
- First came Sharding
- Then came NoSQL

pra-sâmi

35 NoSQL Movement

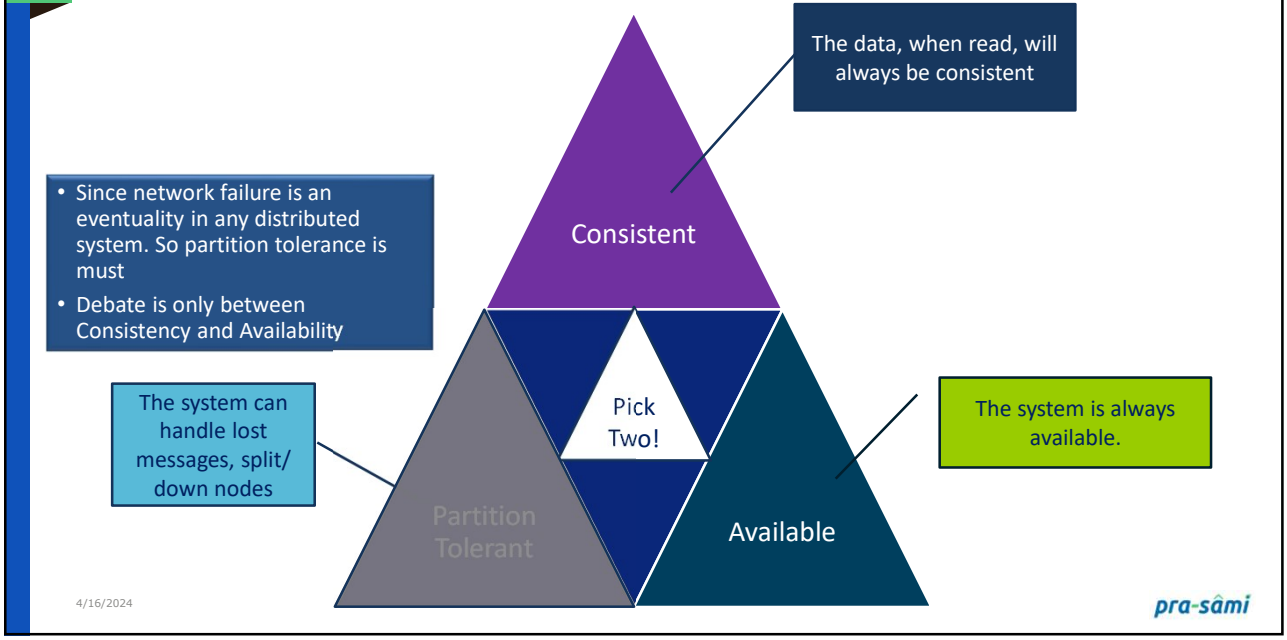


- Grown out of a frustration for traditional RDBMS scalability issues
- Abandons ACID compliance in favor of scalability
- Generally solves for "eventually consistent"
- Far less featured than traditional SQL systems from perspectives of data management, querying, transactions, etc.
- Indexing and other querying optimizations often left to application developers
- Tradeoffs on features are outweighed by scalability concerns for many applications, thus the surge in growth

pra-sâmi

36

CAP Triad

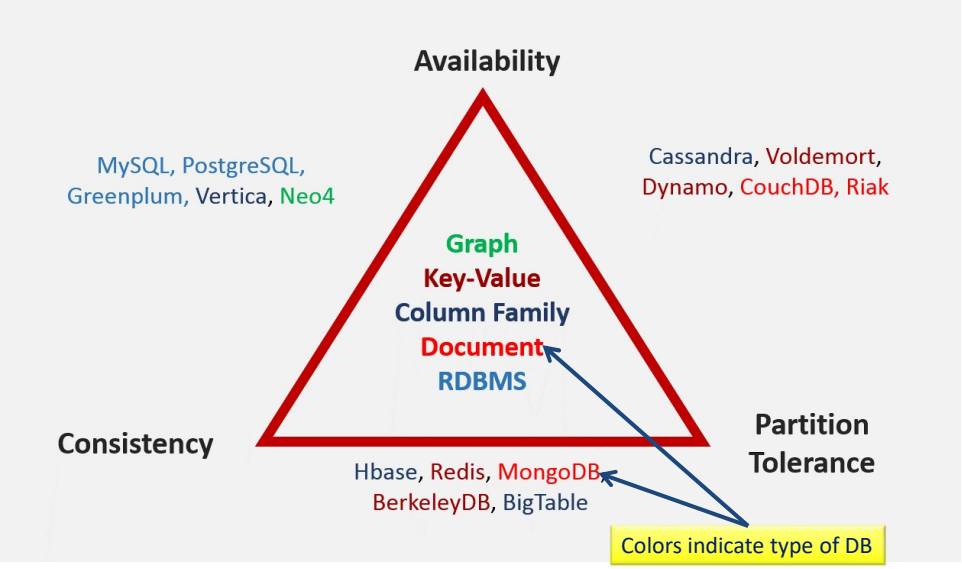


4/16/2024

pra-sâmi

37

Cap Theorem

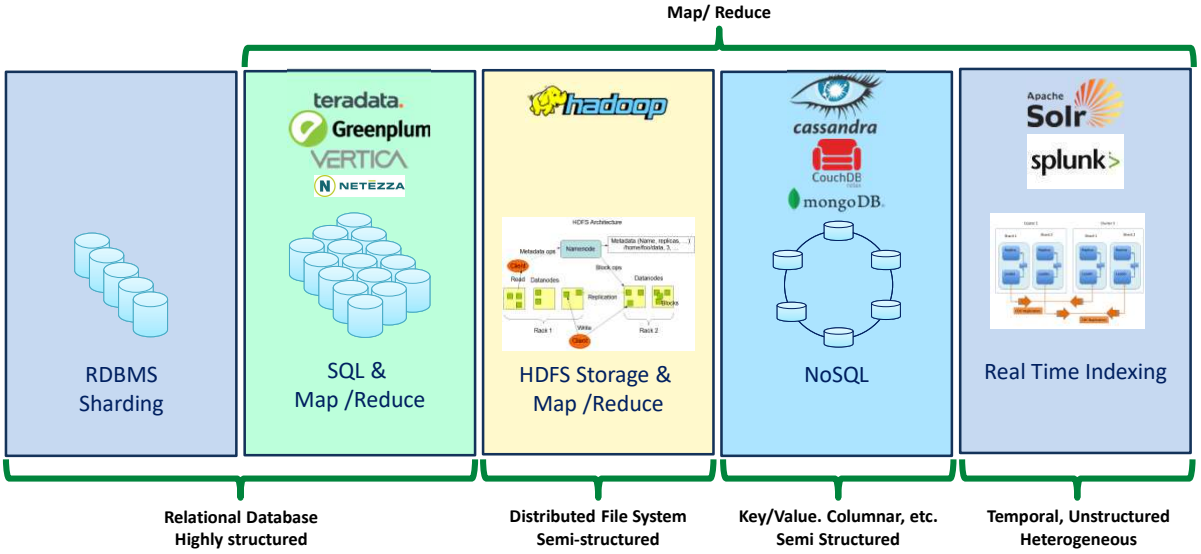


4/16/2024

pra-sâmi

38

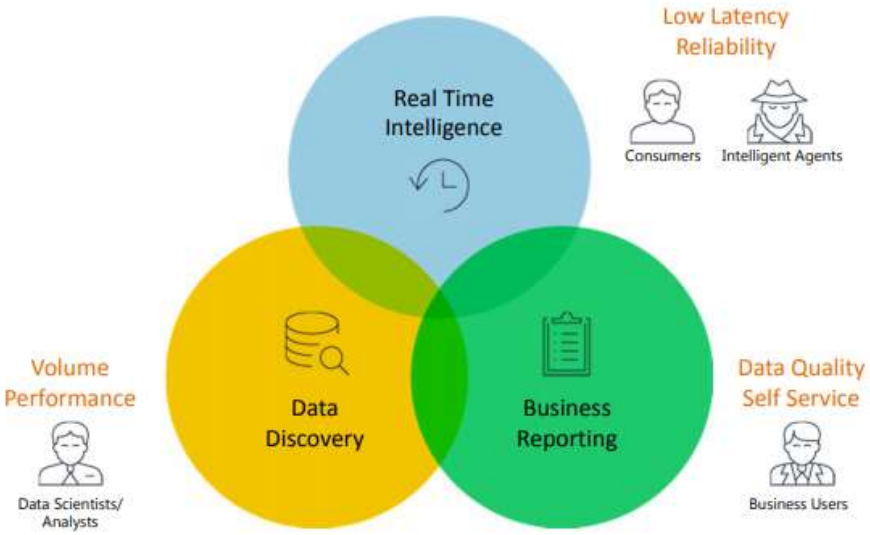
Big Data Technologies



pra-sâmi

39

Big Data Analytics Use Cases



pra-sâmi

40

Big Data Analytics Reference Architectures

Architecture Drivers	Reference Architectures
❑ Volume	❑ Extended Relational
❑ Sources	❑ Non-Relational
❑ Throughput	❑ Hybrid
❑ Latency	
❑ Extensibility	
❑ Data Quality	
❑ Reliability	
❑ Security	
❑ Self-Service	
❑ Cost	

pra-sâmi

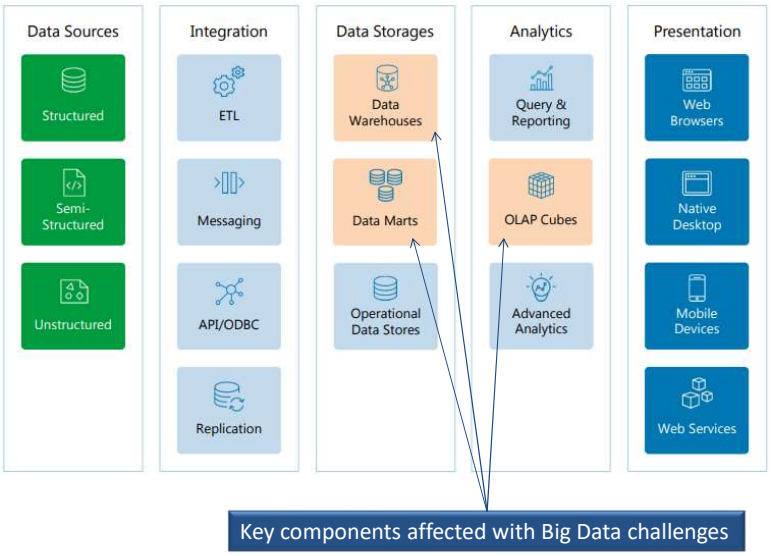
41

Relational Reference Architecture

pra-sâmi

42

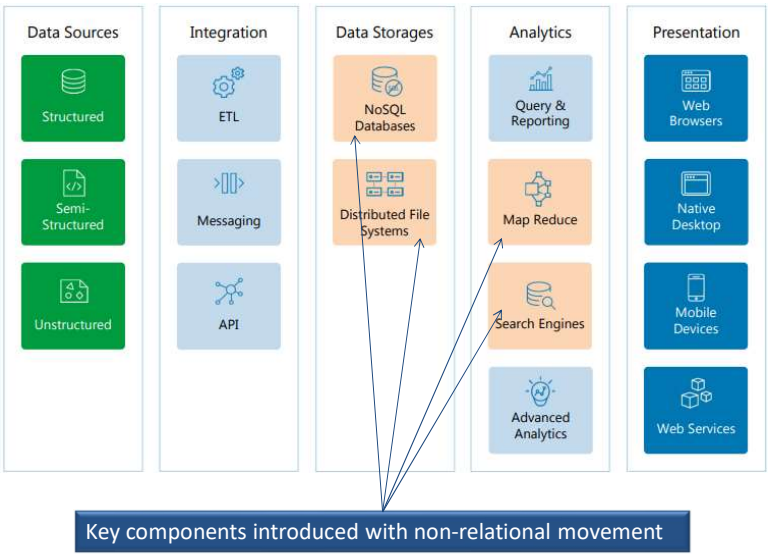
Extended Relational Reference Architecture



pra-sâmi

43

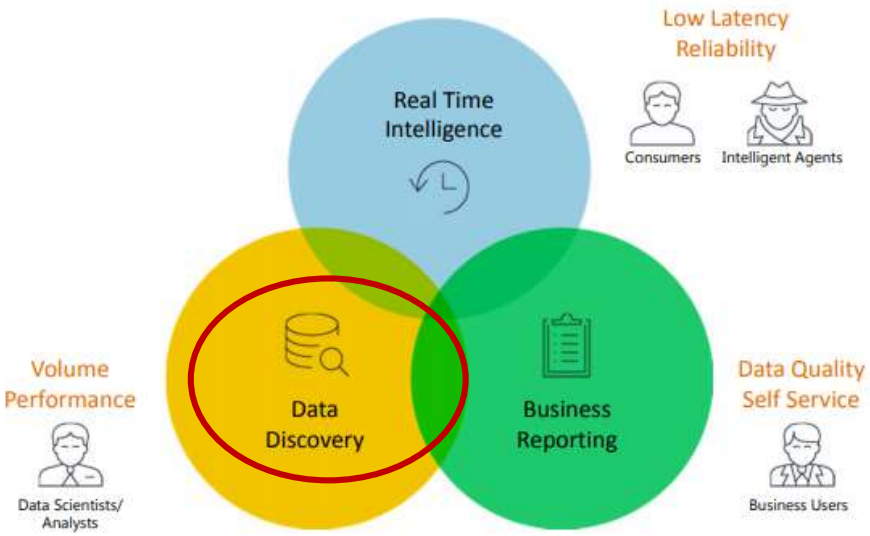
Non-Relational Reference Architecture



pra-sâmi

44

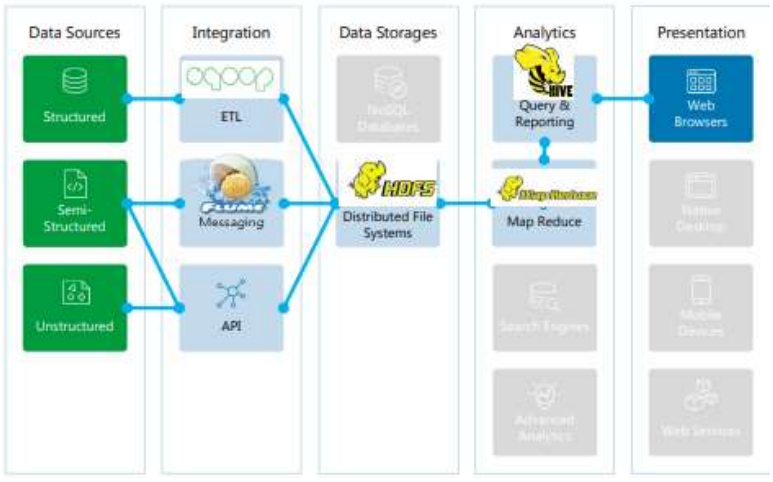
Big Data Analytics Use Cases



pra-sâmi

45

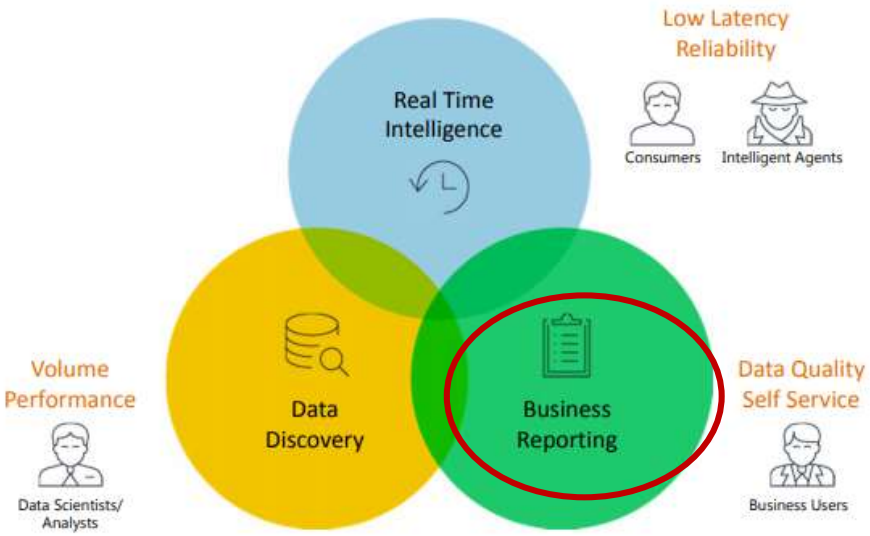
Data Discovery: Non-Relational Architecture



pra-sâmi

46

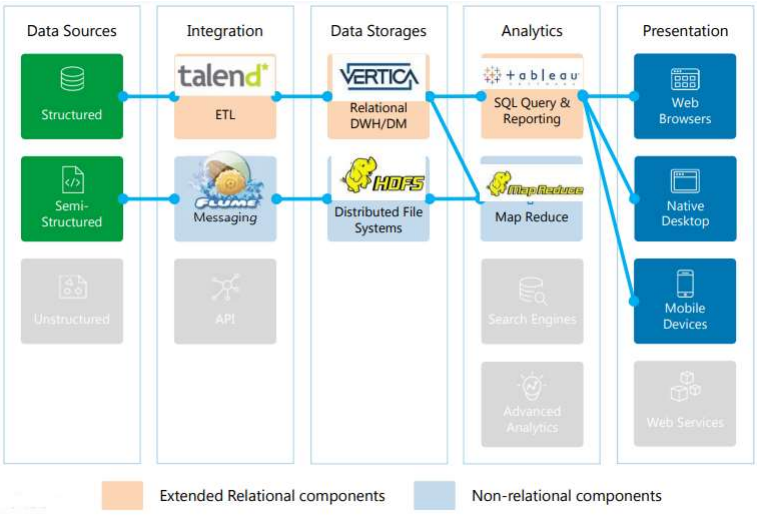
Big Data Analytics Use Cases



pra-sâmi

47

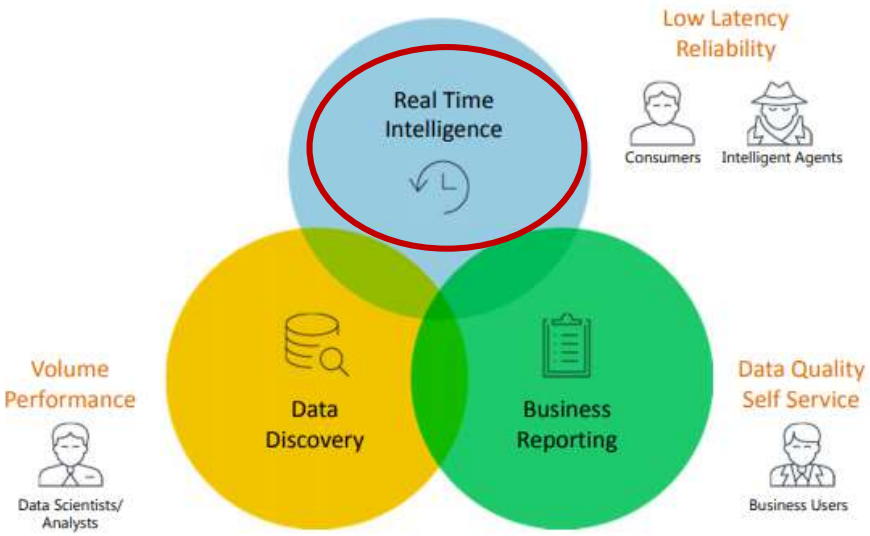
Business Reporting: Hybrid Architecture



pra-sâmi

48

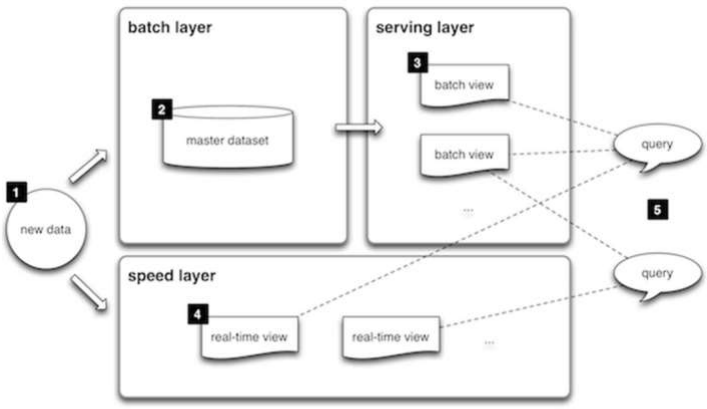
Big Data Analytics Use Cases



pra-sâmi

49

Lambda Architecture



Source: Big Data Principles and Best Practices of Scalable Realtime Data Systems by Nathan Marz

pra-sâmi

50

Apache Cassandra

- Cassandra is a highly scalable semi-structured database
  - Fault Tolerant
  - Scalable
  - Open source
- What Makes Cassandra Different
  - Scales linearly, which is unique for a database product
  - Tunable consistency, allowing speed to be adjusted based on the consistency needs of the application
  - Very fast, especially tuned for eventually consistent, allowing blazingly fast writes
  - Best used for transactional systems which need fast response times and very high scalability
  - Strong commercial backer in DataStax

51

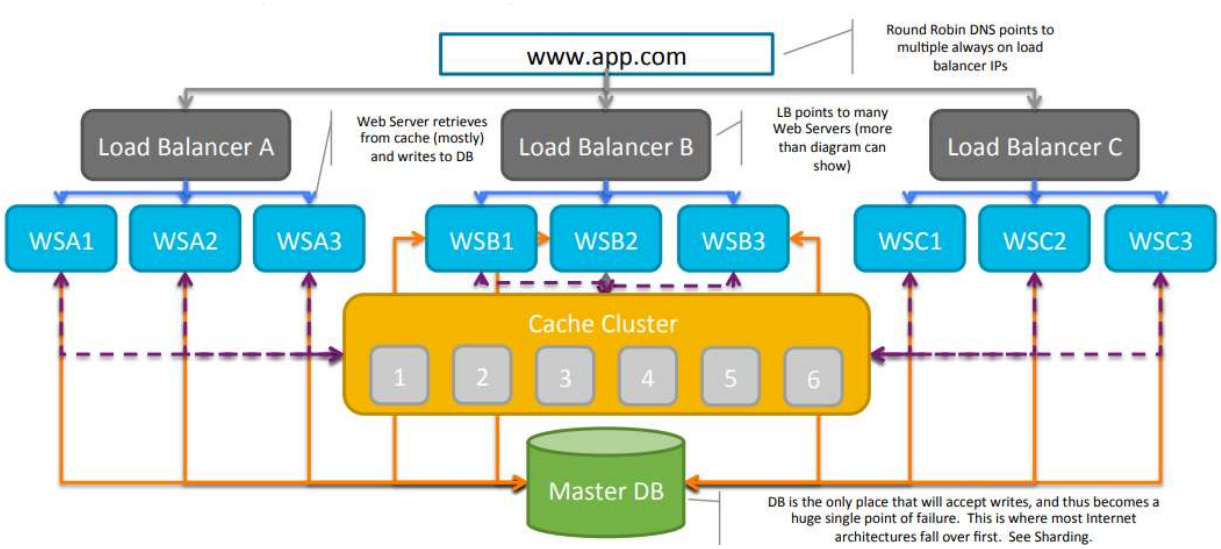
Understanding Cassandra Data Structure

- Nodes are organized into Clusters
- Clusters contain data organized into Keyspaces, similar to a Database Instance or Splunk Index
- Keyspaces can contain multiple “Column Families”, which can be considered analogous to tables
- Columns cannot be required, and additional columns may be added at any time to a given column family
- A column families can have any number of columns or be completely dynamic columns
- Column names can also be used for storing data

Rowkey	Column Family Employee			Column Family Contact	
EmpID	First Name	Middle Initial	Last Name	Phone	Email
1	Joe		Blow	555-1212	joe@...
2	Sara	M	Name		
3	Srinivas			555-1234	
4				555-4321	jsmith@...

52

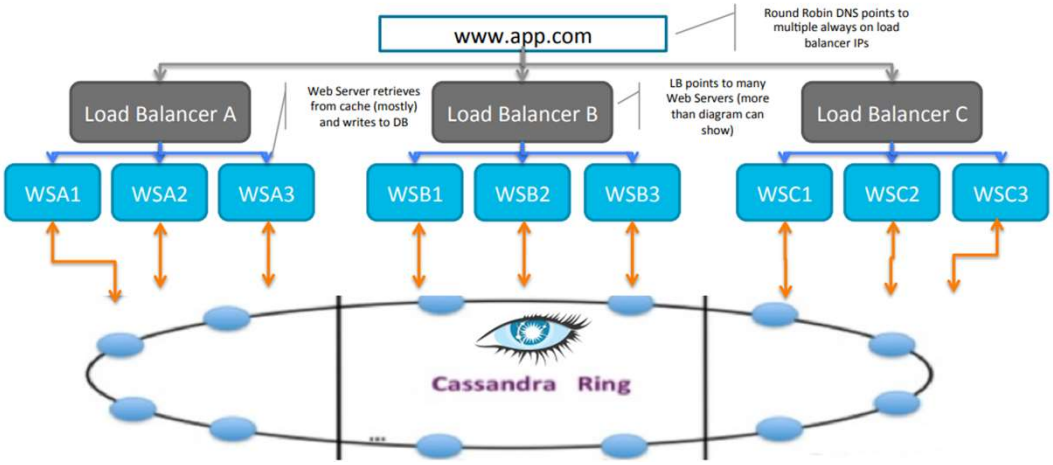
Typical Internet Application Architecture



pra-sâmi

53

Typical Internet Application Architecture

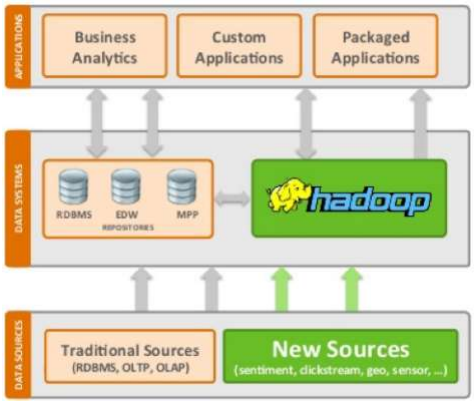


pra-sâmi

54

Movement away from Legacy EDW

- These systems does not work at the scale of data today:
  - The Costs of storage, SQL licensing and required number of humans
- As companies see more value:
  - More and more new data sources are added all the time
  - Most of these do not fit a relational model
- People want a new way to do their own ad hoc analysis and not to rely on a team of EDW resources to make data available in hard reports
- Legacy EDW systems are under strict surveillance and not easy to access or fire ad hoc queries



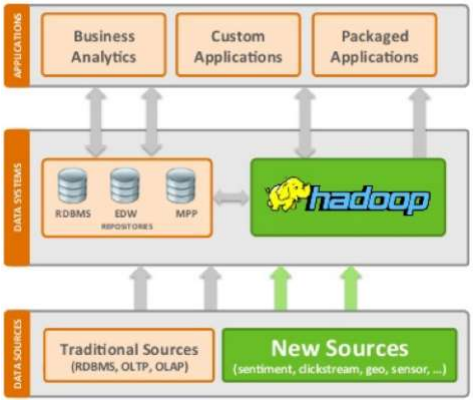
4/16/2024

pra-sâmi

55

Future end of Legacy ETL $$$$

- The Costs of storage, SQL licensing and required number of humans for these systems does not work at the scale of data today
- Changes in incoming formats were throwing away millions of rows
- People are no longer willing to accept the bias of a person in the ETL process to decide how to answer the questions that need to be asked
- People are now starting to see that the golden nuggets in the data have been ETL'ed out for years now that they can get access to all the raw data
- ETL Team should not take a call as to what data to keep and what to throw away.

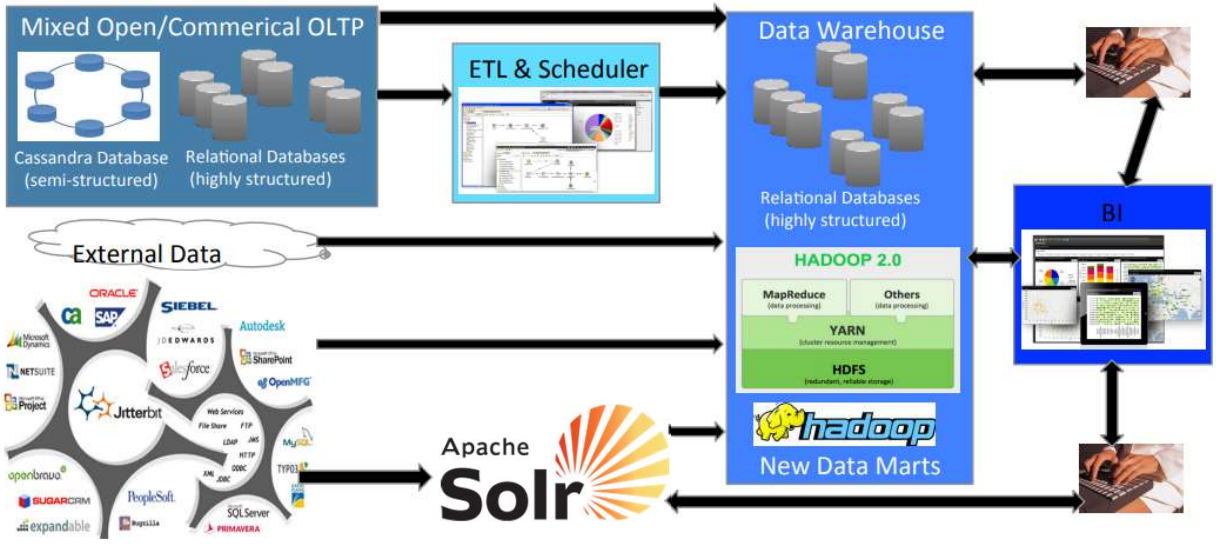


4/16/2024

pra-sâmi

56

End to End View



pra-sâmi

57

Reflect…

- The feature of big data that refers to the quality of the stored data is _____
  - Variety
  - Volume
  - Variability
  - Veracity
- Concerning the Forms of Big Data, which one of these is odd?
  - Structured
  - Unstructured
  - Processed
  - Semi-Structured
- The word ‘Big Data’ was coined in the year
  - 2000
  - 1970
  - 1998
  - 2005
- The word 'Big data' was coined by
  - Roger Mougalas
  - John Philips
  - Simon Woods
  - Martin Green

pra-sâmi

58 Reflect…

- Check below the best answer to "which industries employ the use of so-called "Big Data" in their day to day operations?
  - Weather forecasting
  - Marketing
  - Healthcare
  - All of the above
- Listed below are the three steps that are followed to deploy a Big Data Solution except
  - Data Ingestion
  - Data Processing
  - Data dissemination
  - Data Storage
- The new source of big data that will trigger a Big Data revolution in the years to come is
  - Business transactions
  - Social media
  - Transactional data and sensor data
  - RDBMS
- What makes Big Data Analysis difficult to optimize?
  - Big Data Analysis is not difficult to optimize
  - Both data and cost effective ways to mine data to make business sense out of it
  - The technology to mine data
  - All of the above

59 Reflect…

- Big Data applications benefit the media and entertainment industry by
  - Predicting what the audience wants
  - Ad targeting
  - Scheduling optimization
  - All of the above
- Big data analysis does the following **except:**
  - Collects data
  - Spreads data
  - Organizes data
  - Analyzes data
- The examination of large amounts of data to see what patterns or other useful information can be found is known as
  - Data examination
  - Information analysis
  - Big data analytics
  - Data analysis
- All of the following accurately describe Hadoop, **except** :
  - Open-source
  - Real-time
  - Java-based
  - Distributed computing approach

60

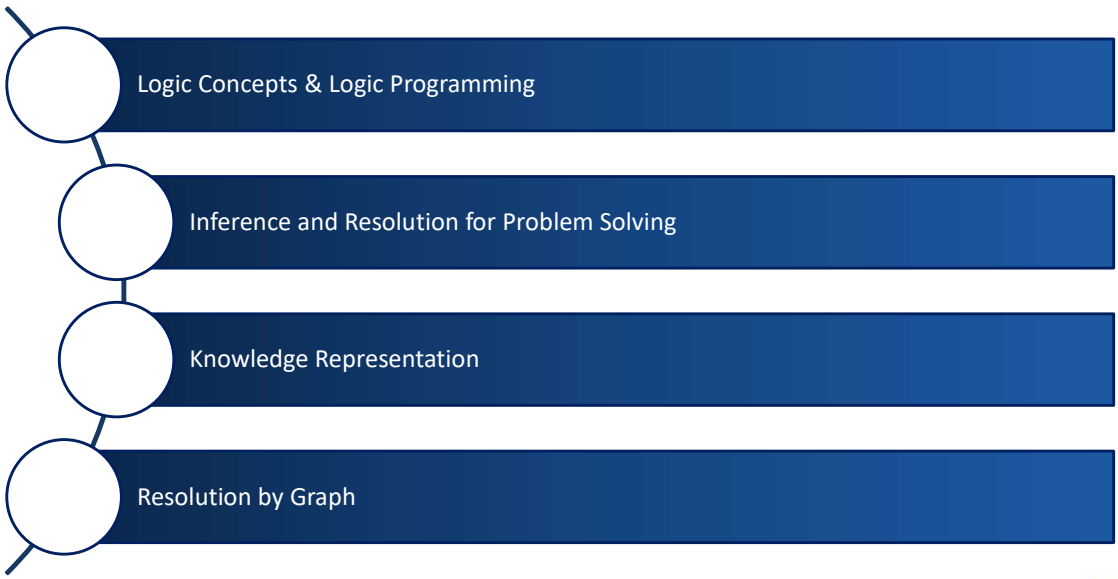
Reflect…

- According to analysts, for what can traditional IT systems provide a foundation when they're integrated with big data technologies like Hadoop?
  - Big data management and data mining
  - Data warehousing and business intelligence
  - Management of Hadoop clusters
  - Collecting and storing unstructured data
- What are the main components of Big Data?
  - MapReduce
  - HDFS
  - YARN
  - All of these
- What are the different features of Big Data Analytics?
  - Open-Source
  - Scalability
  - Data Recovery
  - All the above

pra-sâmi

61

Next Session



pra-sâmi

62

pra-sâmi

ADDITIONAL MATERIAL

pra-sâmi

64

Total Chaos

pra-sâmi

65

Chaos

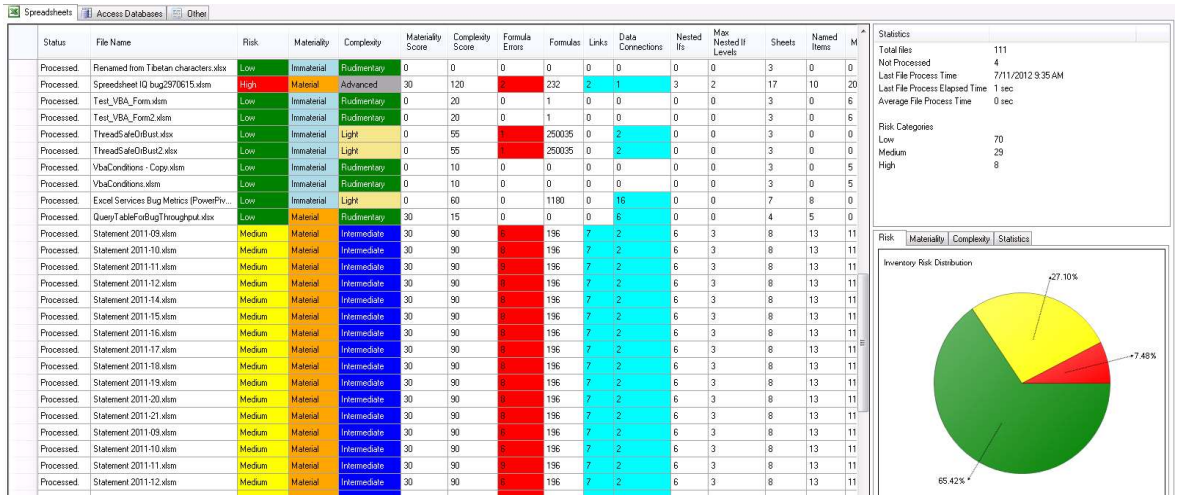
- We have over 10,000 Excel spreadsheets in the organisation.
- I am going to ban Excel

- How many time we have heard this type of statements?

pra-sâmi

66

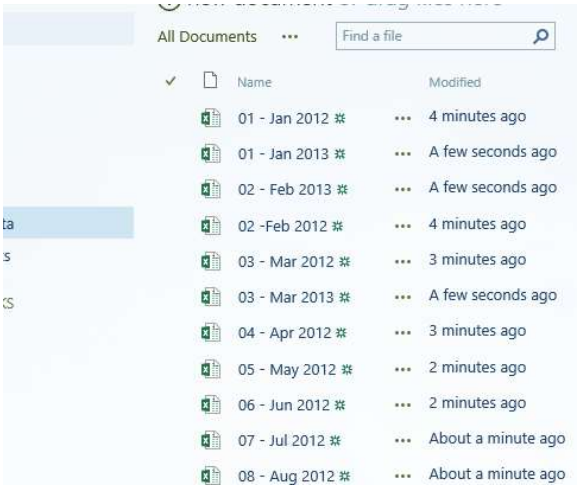
Discovery

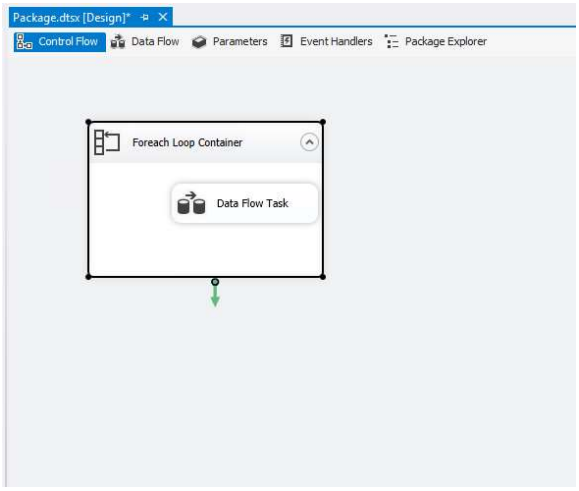


pra-sâmi

67

Data Source

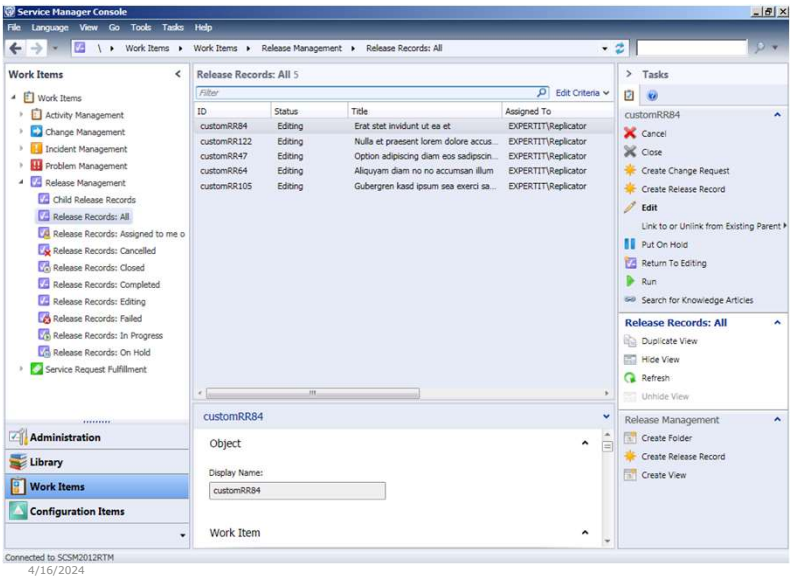




pra-sâmi

68

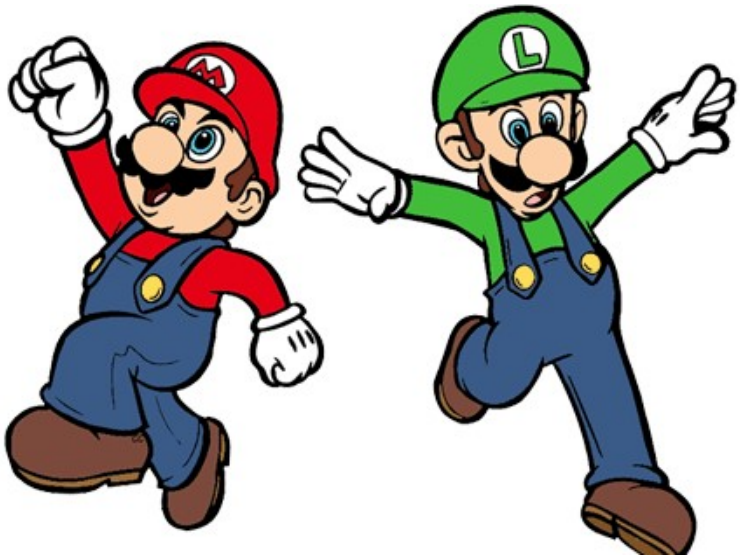
Change Control



pra-sâmi

69

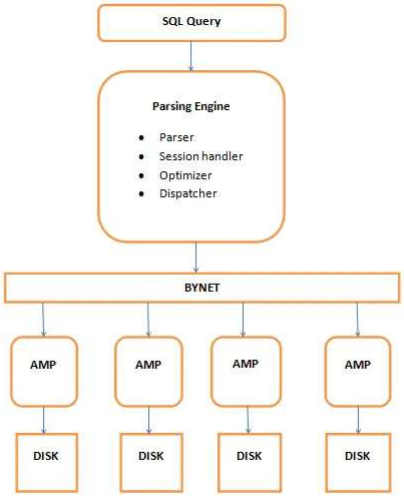
Success



pra-sâmi

70

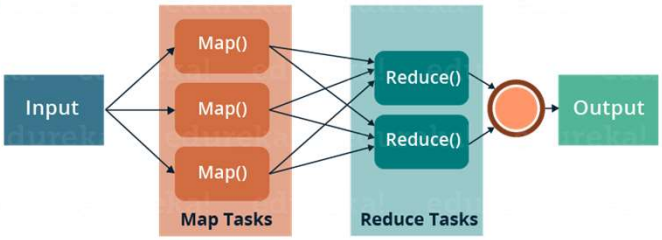
Teradata Architecture

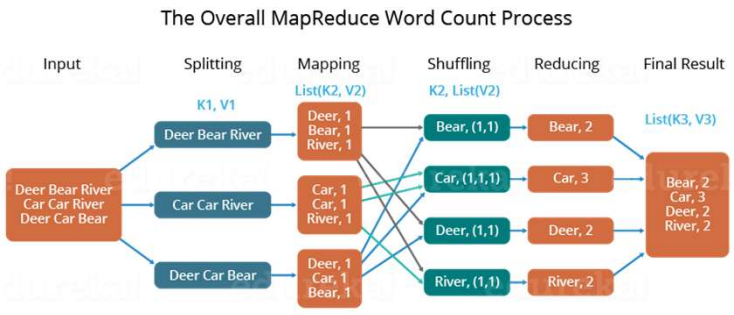


pra-sâmi

71

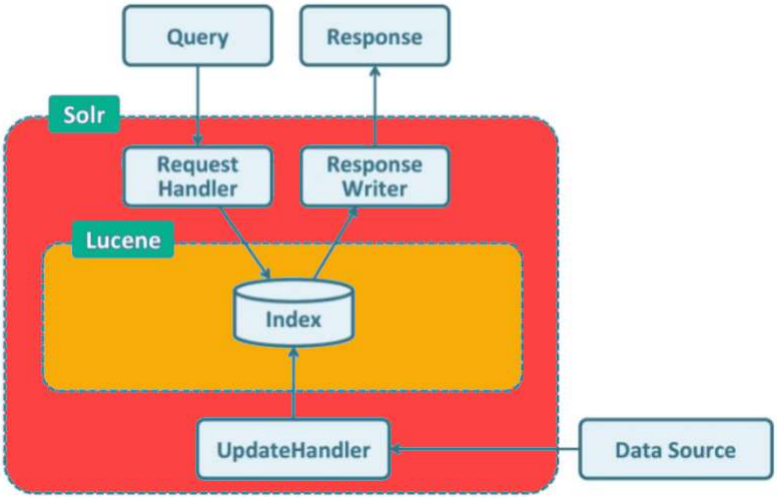
Map Reduce





pra-sâmi

72 Apache Solr



pra-sâmi

73 In Memory Data Grid

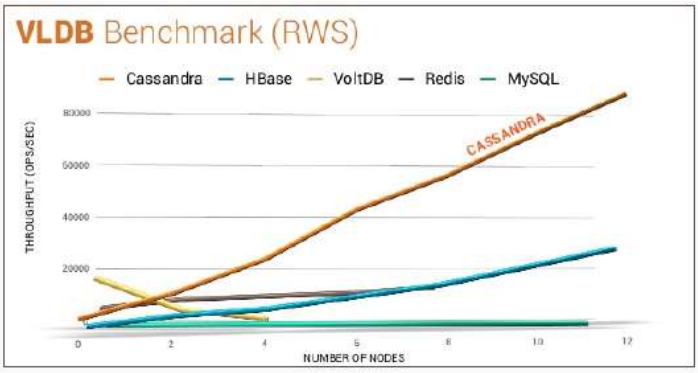
- Distributed in-memory database
  - Scale out (Horizontal scaling)
- Pros
  - Scale on write/read
  - Fits to event driven (CEP style) , ad-hoc query model
- Cons
  - Cost of memory vs. disk
  - Memory capacity is limited
- Apache Spark
  - Generality
    - Combine SQL, streaming, complex analytics.
  - Runs Everywhere
    - Spark runs on Hadoop, Mesos, standalone, or in the cloud. It can access diverse data sources (HDFS, Cassandra, HBase, and S3)
  - Ease of Use
    - Write applications quickly in Java, Scala, Python, R.

*Generally accompanied with Kafka and Zookeeper

pra-sâmi

74

VLDB Benchmark (RWS)



pra-sâmi