



BIG DATA

Introduction : Characteristics and Architecture

K V Subramaniam

Computer Science and Engineering
subramaniamkv@pes.edu

BIG DATA

Course Introduction Contd...



1. Big Data – Characteristics
2. Data Architecture Design
3. Data Format/Types
4. Big Data Platforms
5. Case Study - Google

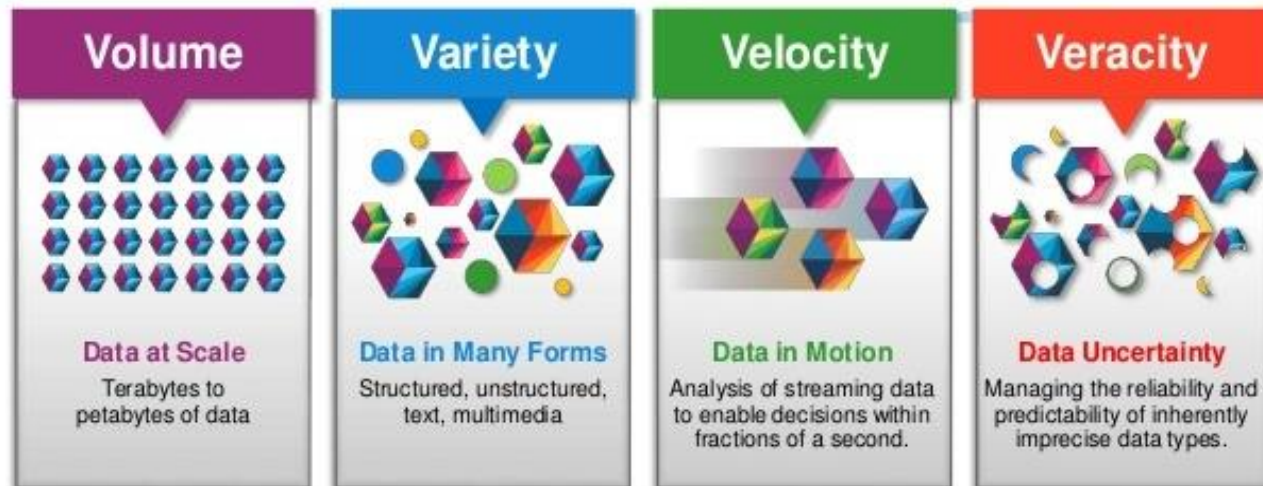
BIG DATA

Characteristics – Quick Summary



Big Data - 4Vs

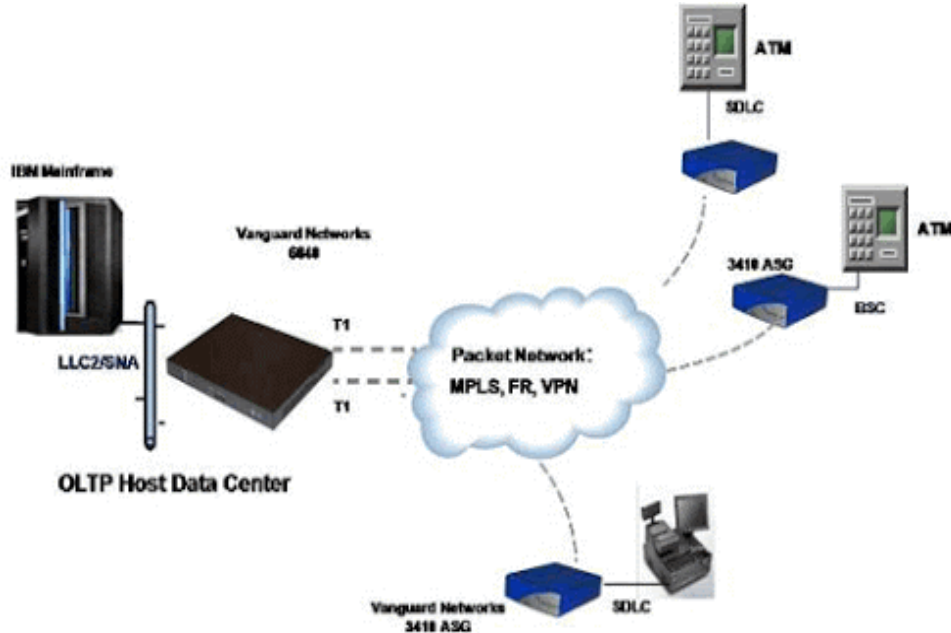
* IBM's 4Vs of Big Data:



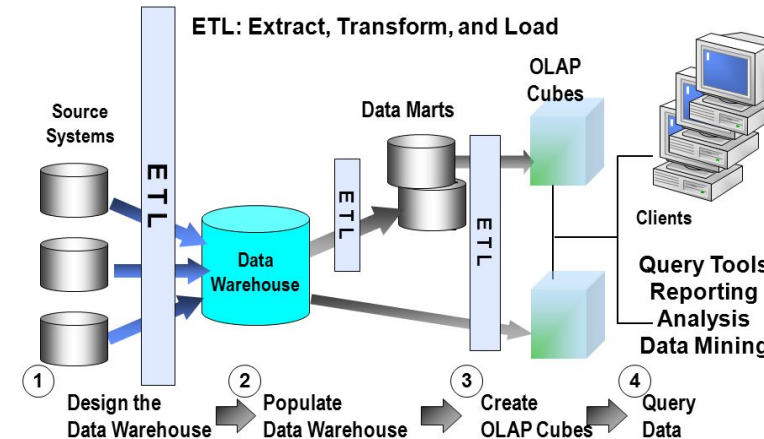
Big Data: Characteristics - Volume

BIG DATA

Old Style Data



The Data Warehouse/Bi Architecture & Process



© Minder Chen, 2004-2014

DW & BI - 13

<http://www.techbridge.solutions/home/solutions/solutions-ibm-applications/oltp-host-concentrator-solution/>

<https://data-flair.training/blogs/business-intelligence-and-data-warehousing/>

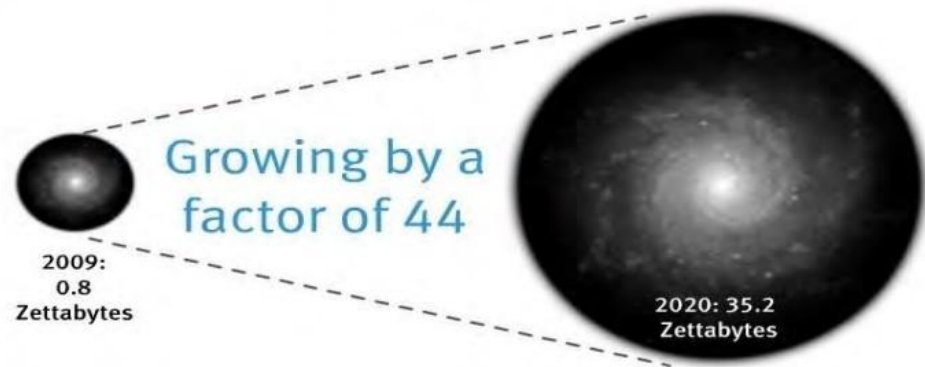
- Fixed schema and format
 - Clean data
 - Consistent
 - Predictable data rates
- Elaborate ETL procedure (Not real time; once a day at best)
 - Output mostly reports (Queries run in batches; take hours)

BIG DATA

Characteristics - Volume

- Old Style Data vs Big Data
1-Scale (Volume)
 - 44x increase from 2009 2020
 - From 0.8 zetta bytes to 35zb
 - Data volume is increasing exponentially

The Digital Universe 2009 to 2020



Equivalent to a stack of DVDs in 2009 reaching to the moon and back, now reaching halfway to Mars by 2020

BIG DATA

Characteristics - Volume



PES
UNIVERSITY
ONLINE

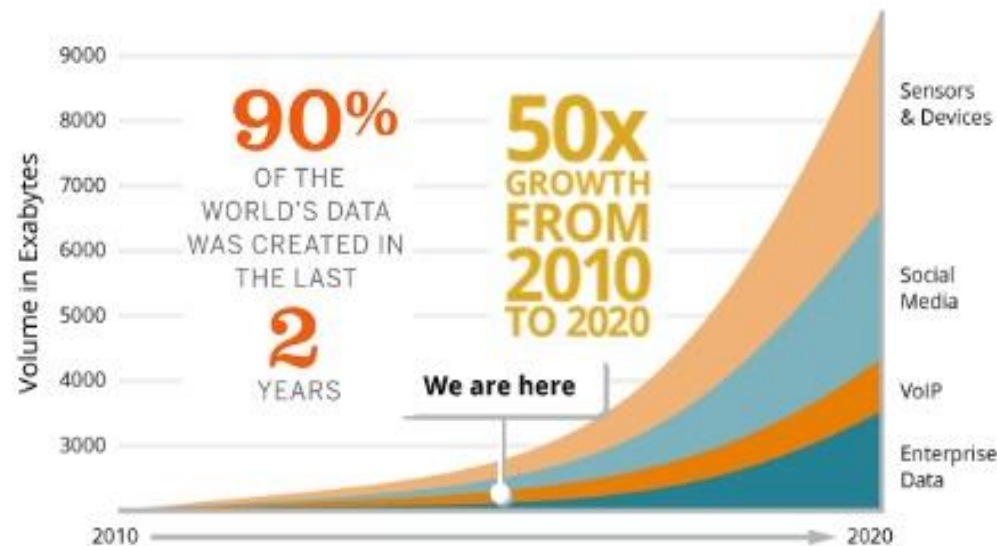
- Why is the volume so high? Who's Generating it?

CONTEXT: WHAT'S BIG DATA?

7

BIG IN GROWTH, TOO.

1 exabyte (EB) = 1,000,000,000,000,000 bytes



Old Style Data

Big Data

Enterprises
only

Everybody

BIG DATA

Characteristics - Volume

- How this VOLUME gets collected?

Old Style Data	Big Data
Manual entry	Automated

- Financial Services

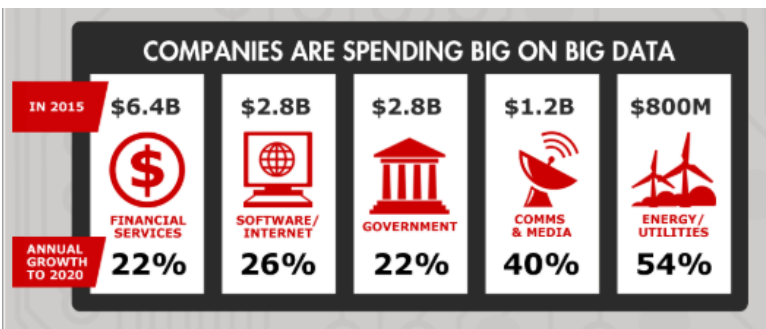
- Stock trading, banking, financial services are all computerized (In India – no paper shares any more)
- Every transaction is recorded

- Energy / Utilities

- Can put sensors in every home
- Smart grid

- Comms & media

- Internet services



<https://www.forbes.com/sites/louiscolumnbus/2014/06/24/roundup-of-analytics-big-data-business-intelligence-forecasts-and-market-estimates-2014/#26c06d11388e>

Big Data: Characteristics - Variety

BIG DATA

Characteristics - Variety

- Why is analyzing Big Data *complex*?
- Various formats, types, and structures
- Text, numerical, images, audio, video, sequences, time series, social media data, multi-dimensional arrays,
- Static data vs. streaming data
- A single application can be generating/collecting many types of data

Old Style Data	Big Data
Fixed format / schema	Integrate Twitter, Maps, Facebook

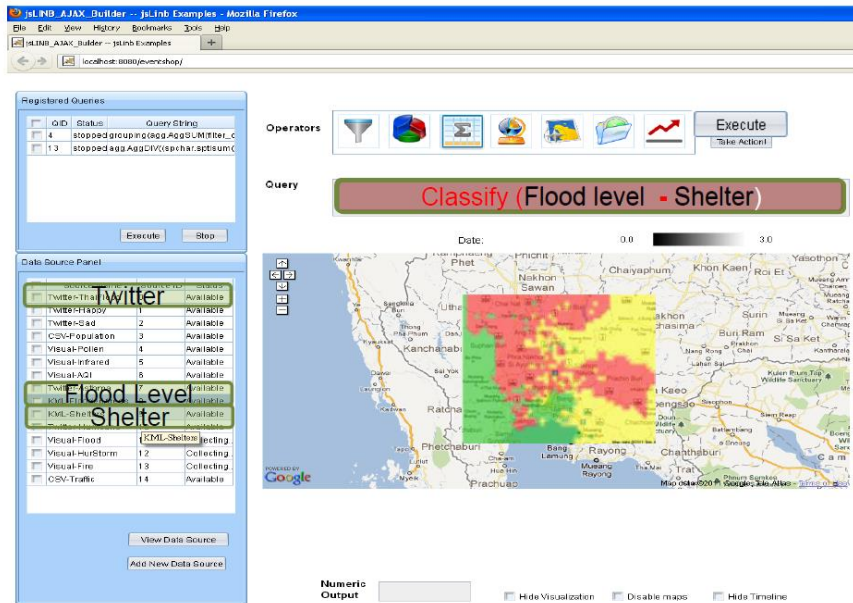
To extract knowledge all these types of data need to be linked together



BIG DATA

Characteristics - Variety

EventShop: Thai Floods



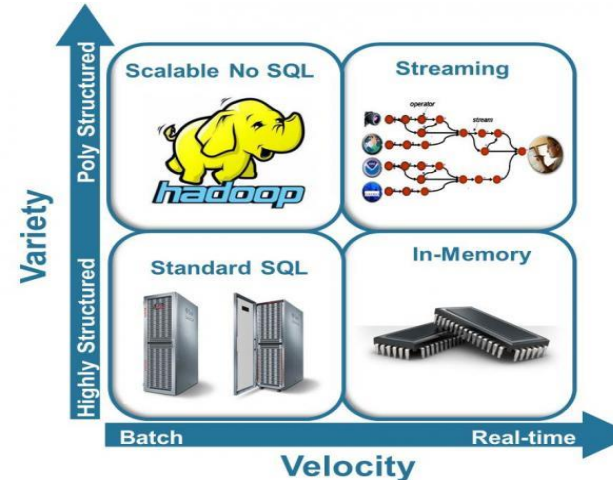
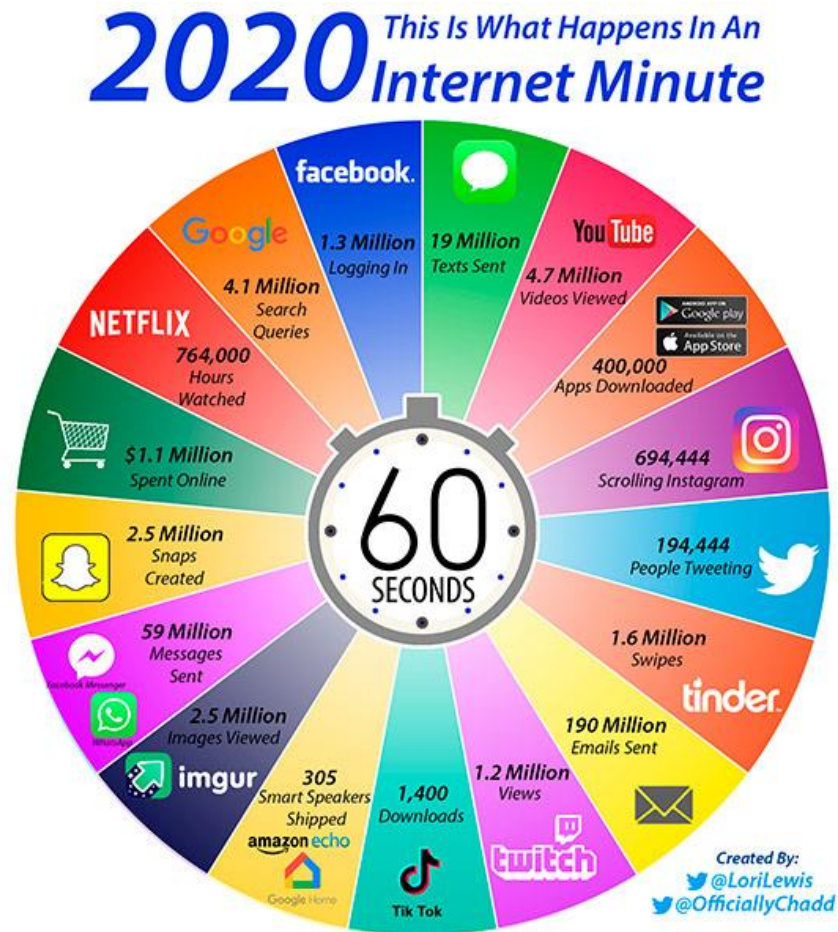
SHE SAYS

How Twitter transformed itself during
#ChennaiFloods for me

Big Data: Characteristics - Velocity

BIG DATA

Characteristics - Velocity



Source: Forrester Webinar: Big Data: Gold Rush or Illusion?, Sept 19, 2013

<https://go.forrester.com/blogs/14-05-27-boost-your-digital-intelligence-with-big-data/>

BIG DATA

Characteristics - Velocity

Nokia launches real-time mobile network analytics platform

Staff writer | July 28, 2016 | Telco Analytics Asia

<https://www.nokia.com/about-us/news/releases/2016/06/30/nokia-real-time-mobile-network-analytics-provides-instant-correlation-of-network-wide-data-and-its-impact-on-customer-experience/>

The analytics link the performance of applications and devices to network issues in real-time, effectively making every mobile device part of a network test bed.

→ This enables operators to pinpoint potential causes of service degradation much more rapidly than they can today since they no longer need to consult a myriad of tools and correlate the data from them manually. It also offers engineering teams a proactive way to understand **over the top (OTT) application** impacts on the network and optimize opportunities.

Old Style Data	Big Data
Input data at night	Real-time input
Output daily report	Real-time response

Big Data: Characteristics - Veracity

Veracity

refers to the messiness or trustworthiness of the data.

Accuracy cannot be controlled due to

- Hashtags
- Typos
- Abbreviations

Need to work with such data

Traditional: manually entered, fixed fields, less chance of error

Old Style Data	Big Data
Clean data	Inconsistent data

- How people get multiple PAN cards
- Slight changes in name and address
P E S U n i v e r s i t y , 1 0 0 f t R o a d ...
PES university, BSK III stage...
Pes University, Dwaraka Nagar...
- Names and addresses d o n ' t m a t c h
- But postman / courier will deliver to correct place!!!

Big Data: Data Architecture Design

BIG DATA

Data Formats / Types

Structured Data



0.103	0.176	0.387	0.300	0.379
0.333	0.384	0.564	0.587	0.857
0.421	0.309	0.654	0.729	0.228
0.266	0.750	1.056	0.936	0.911
0.225	0.326	0.643	0.337	0.721
0.187	0.586	0.529	0.340	0.829
0.153	0.485	0.560	0.428	0.628

Databases

Unstructured Data



Documents, Tweets,
Videos

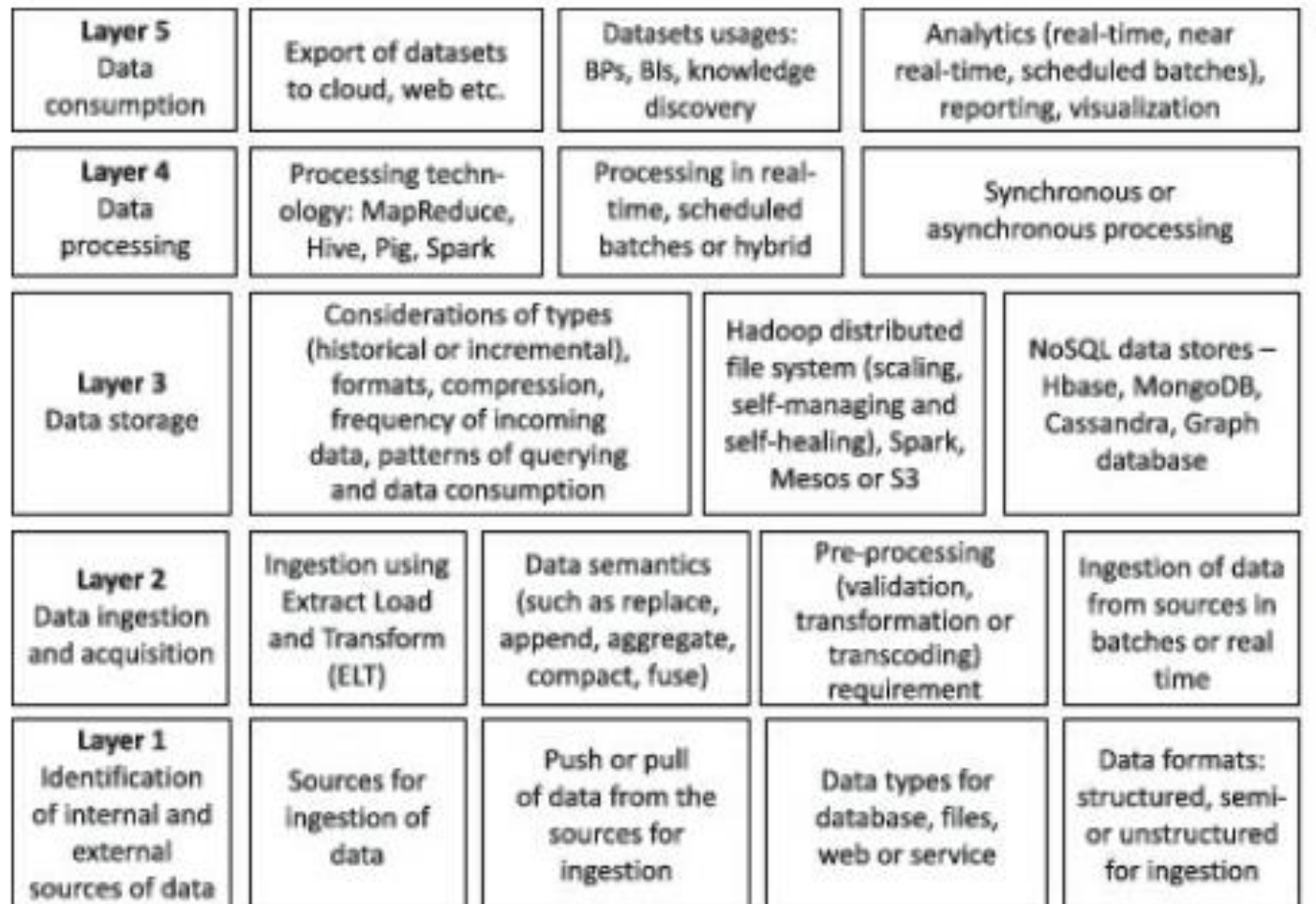
Semi Structured Data



Emails, XML

BIG DATA

Data Architecture Design

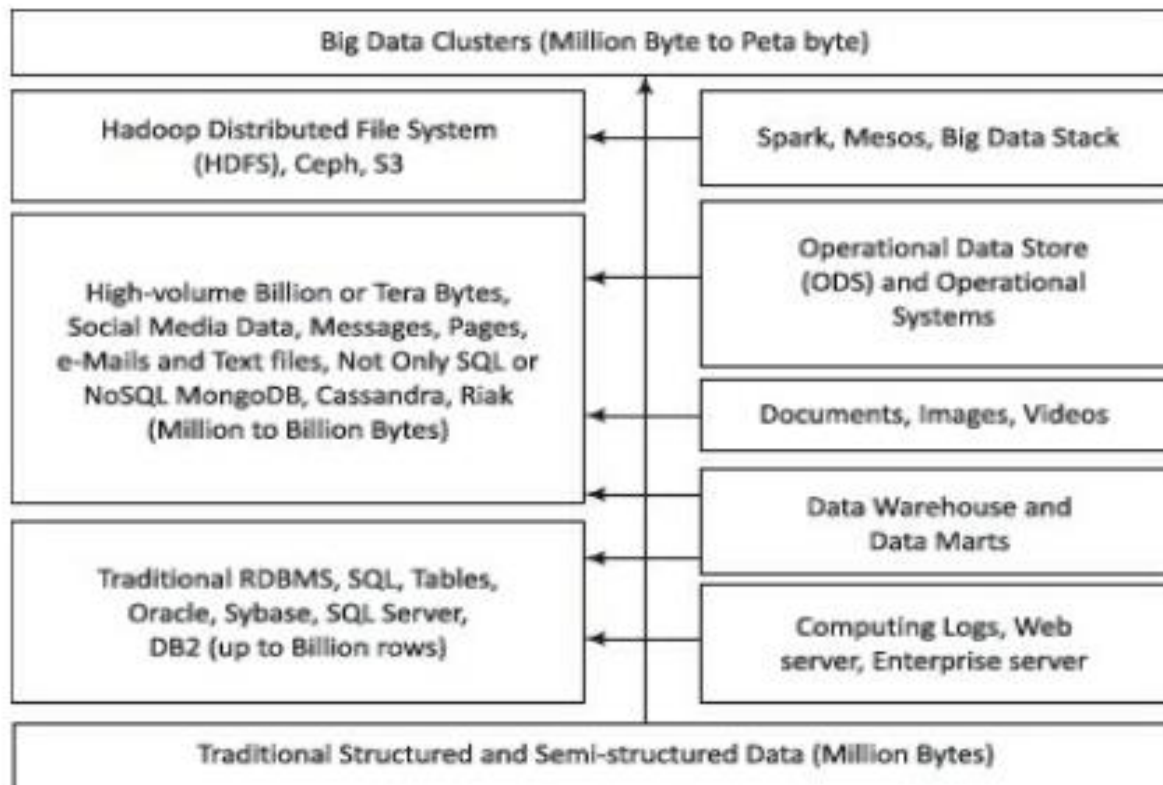


T1: Fig 1.2 – Logical layers in data processing architecture and their functions.

BIG DATA

Data Formats / Types : Big Data Storage

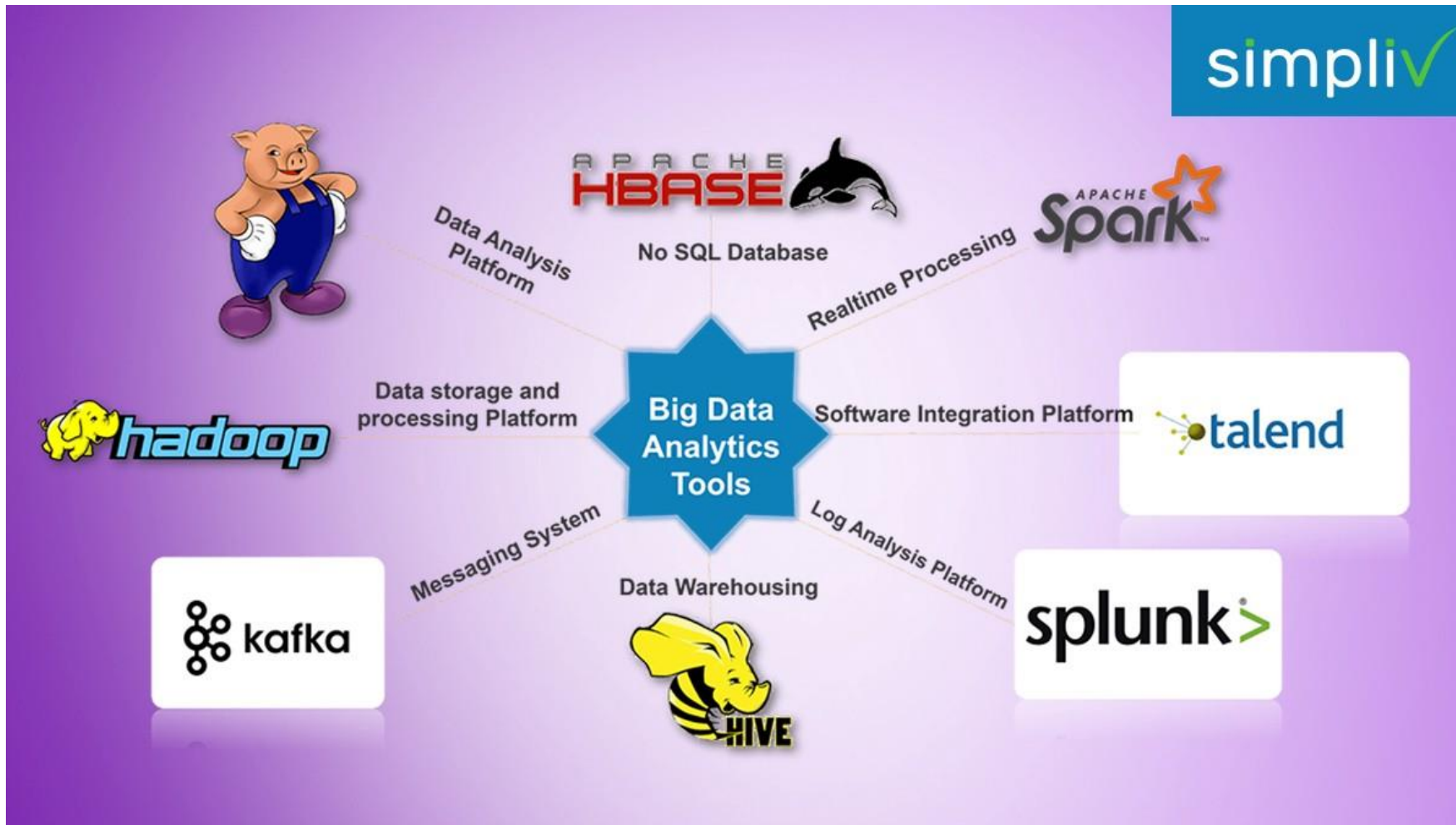
- Data storage for traditional and Big Data



T1: Fig 1.7 – Big data storage plan – RDBMS and NoSQL together

BIG DATA

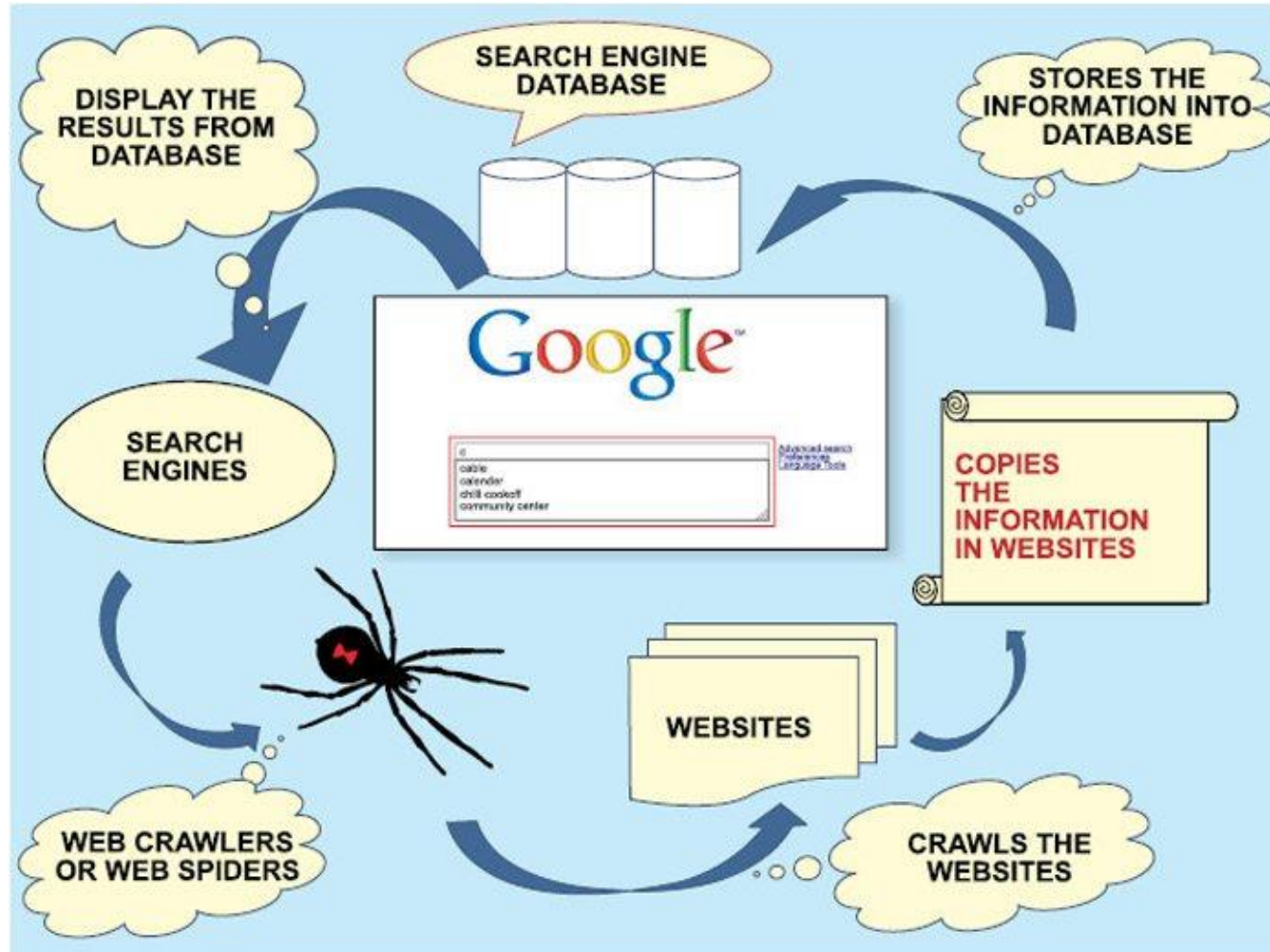
Big Data Platforms



Big Data: Case Study – Google Search

BIG DATA

Case Study: How Google Search works?





- Google initial implementation
 - 24 M web pages, 322 M links
 - 5 days to compute **Page Rank**
 - Page rank is proportional to the popularity of the page
 - If more links point to a page, that page will be more popular
- Page Rank – Treats the web as a graph
 - User -
 - Starts at a random page
 - Takes a random link
 - Page Rank Assumptions –
 - The more popular a page - The better is its quality



<https://en.wikipedia.org/wiki/PageRank>



THANK YOU

K V Subramaniam

Computer Science and Engineering

subramaniamkv@pes.edu

+91 80 6666 3333 Extn 877