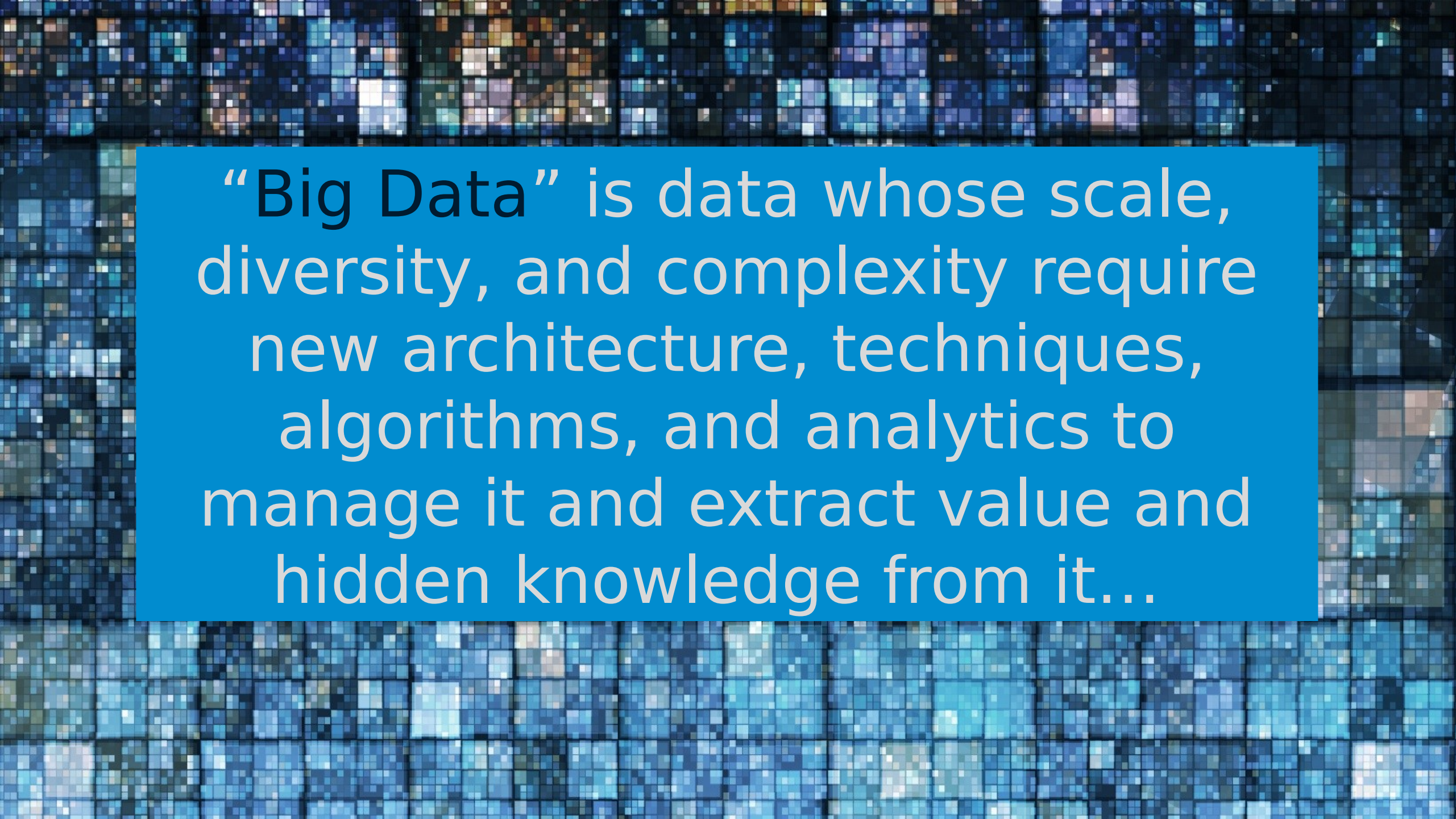

Introduction To Big Data



Outline

- What is "Big Data"
- Big Data Foundational Technologies
- The structure of data projects
- Data visualization



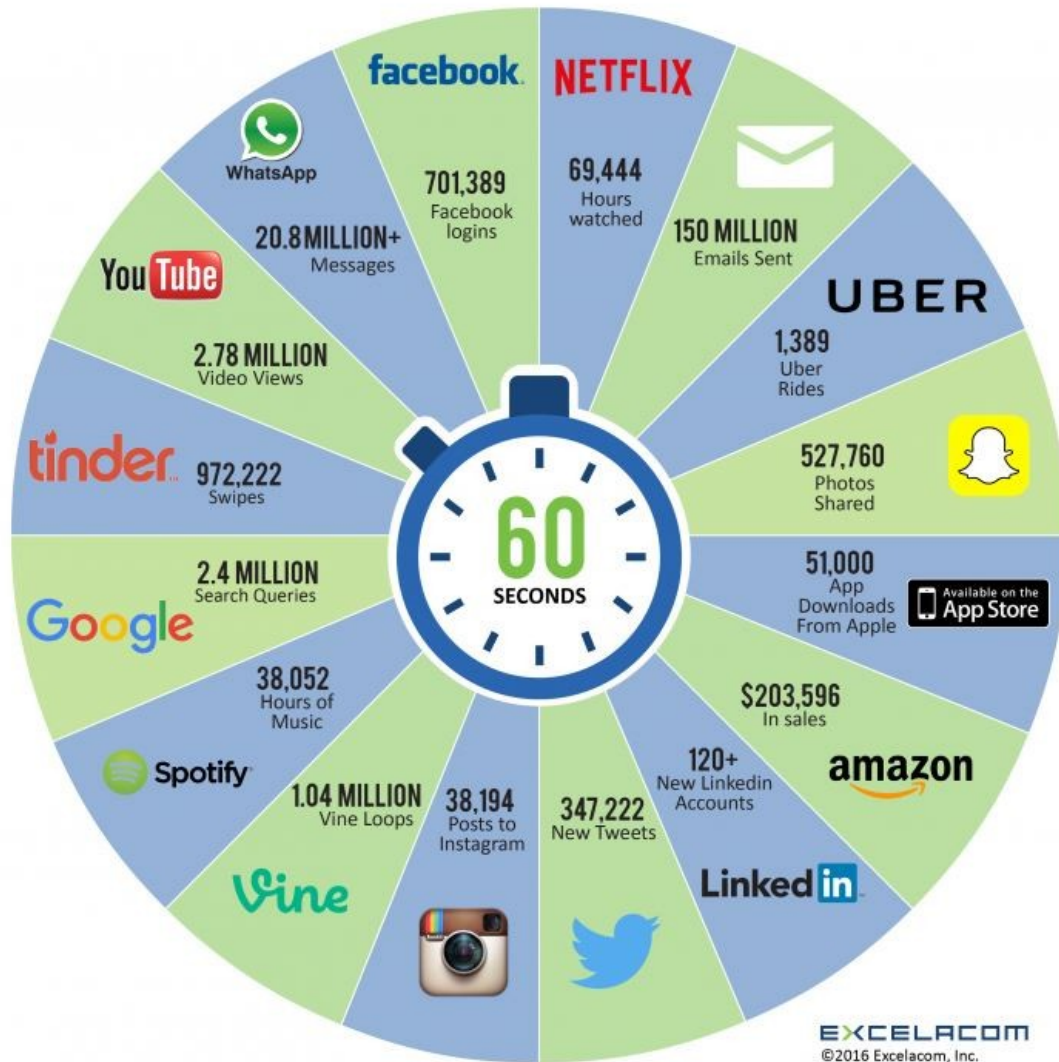
“Big Data” is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it...

THE BIG DEAL ABOUT BIG DATA

Goldman
Sachs

Asset

2016 What happens in an INTERNET MINUTE?



The Scale of Big Data

90% Of today's data has been created in the last two years

Every day we create 2.5 quintillion bytes of data or enough to fill 10 million Blu-ray discs

Most companies in the US have over 100 terabytes (100,000 gigabytes) of data stored

40 zettabytes (40 trillion gigabytes) of data will be created by 2020, an increase of 300 times from 2005, and the equivalent of 5,200 gigabytes of data for every man, woman and child on Earth

2019 *This Is What Happens In An Internet Minute*



Big Data is Constantly Generated and Consumed

From smartphones to social media posts, people create and consume data every second of every day. Big Data is a popular term used to describe the growth and availability of this data. It also refers to the technologies and analytics that collect, manage and extract useful insights.

In only one hour, we generate...



21.6 million

Tweets

Source: Twitter



8.5 billion

Emails

Source: The Radicati Group



34,200

Websites

Source: Forbes



144 million

Google Searches

Source: Google

FORTNITE

A Fortnite game scene featuring several characters in a mountainous landscape. In the foreground, a large tomato character with a mustache and a green leafy top is holding a large black assault rifle. In the background, three other characters are running: one in a blue and white outfit, one in a purple and black outfit, and one in a red and black outfit. The sky is blue with some clouds.

Realtime Ingestion

- Fortnite processes 92 million events a minute and sees its data grow 2 petabytes a month
- Up to 5000 kinesis shards

Data Warehouse

- 14 Petabytes
- 2 PB/month growth rate

"We use the data for everything from ARPU to game analysis and improvements"

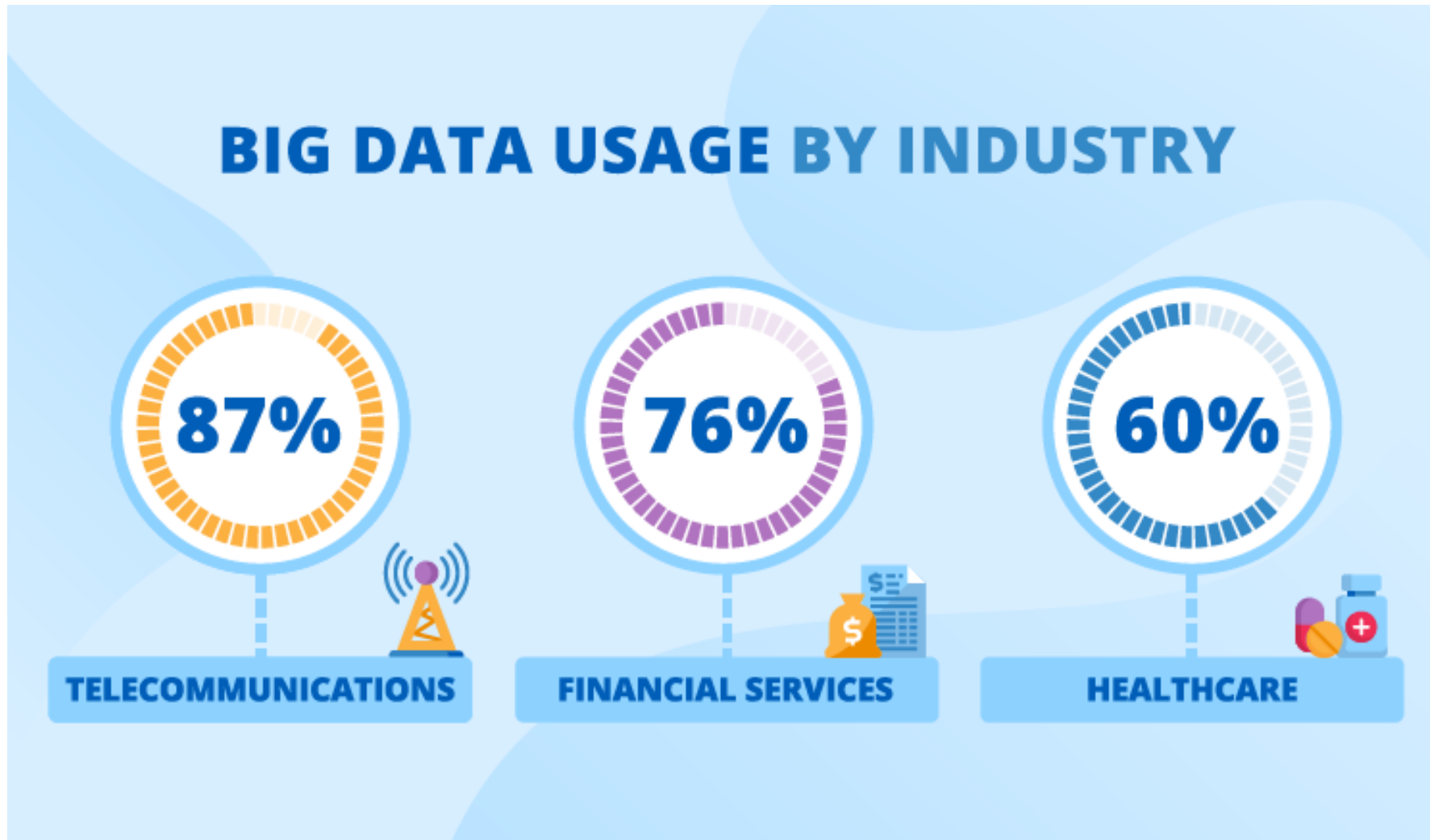
<https://www.zdnet.com/article/how-fortnite-approaches-analytics-cloud-to-analyze-petabytes-of-game-data/>

Figure 1. Magic Quadrant for Data Management Solutions for Analytics



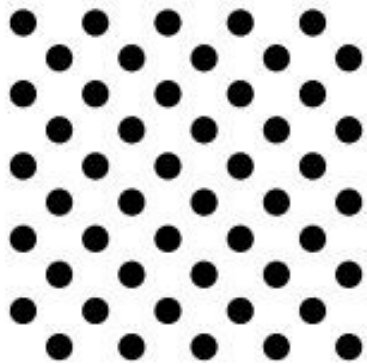
Source: Gartner (January 2019)

Three Industries Most Active in Big Data



The Challenges in Big Data: The four V's

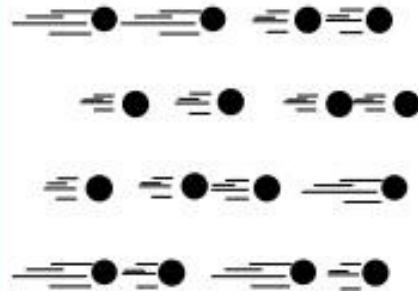
Volume



Data at Rest

Terabytes to exabytes of existing data to process

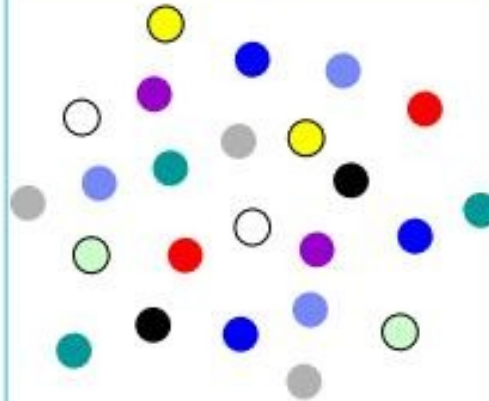
Velocity



Data in Motion

Streaming data, milliseconds to seconds to respond

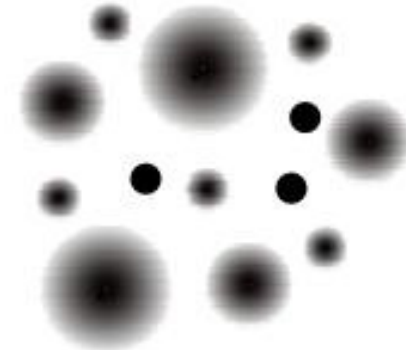
Variety



Data in Many Forms

Structured, unstructured, text, multimedia

Veracity*



Data in Doubt

Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Measuring Data

Unit	Value	Size
bit (b)	0 or 1	1/8 of a byte
byte (B)	8 bits	1 byte
kilobyte (KB)	1000^1 bytes	1,000 bytes
megabyte (MB)	1000^2 bytes	1,000,000 bytes
gigabyte (GB)	1000^3 bytes	1,000,000.000 bytes
terabyte (TB)	1000^4 bytes	1,000,000,000,000 bytes
petabyte (PB)	1000^5 bytes	1,000,000,000,000,000 bytes
exabyte (EB)	1000^6 bytes	1,000,000,000,000,000,000 bytes
zettabyte (ZB)	1000^7 bytes	1,000,000,000,000,000,000,000 bytes
yottabyte (YB)	1000^8 bytes	1,000,000,000,000,000,000,000,000 bytes

Challenges in Analysis of Big Data

1

Inconsistent, incomplete ,
unavailable, poor quality or invalid
data

2

Poor analysis/analytics leading to
erroneous correlations/conclusions

3

Unstructured Data

4

Data privacy/security, employee
confidentiality, personally
identifiable information

5

Corporate data lives in silos

6

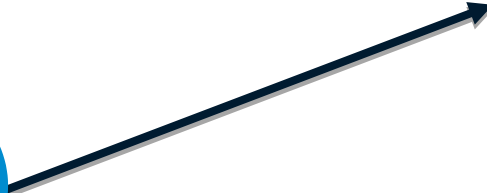
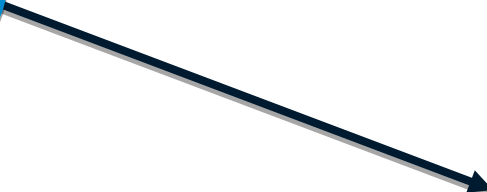
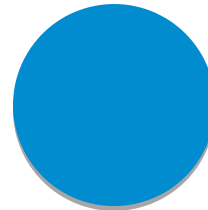
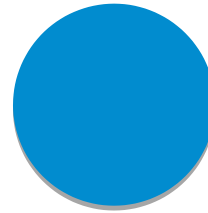
The data does not explain why

Sources of Data

- Paper
- Spreadsheet
- Proprietary format
- Open format
- Available through API
- Available export from application
- Data feeds
- Communications
- Other

Internal Data

External Data



Data Warehousing

- Data is integrated from multiple systems.
- For example provide a full view of a customer:
 - Sales activity
 - Delinquent invoices
 - Support/help requests
- Focus is on reading the information and creating analysis
- Data modelling and ETL process consume most of the time and effort in setting up a data warehouse



Data Lake

- Repository for analyzing large quantities of disparate sources of data in its native or raw format
- Reduce up-front effort by ingesting data in any format without requiring a schema initially
- Make acquiring new data easy, so it can be available for data science & analysis quickly
- Store large volume of multi-structured data in its native format
- <https://www.mongodb.com/databases/data-lake-vs-data-warehouse-vs-database>



Data Lake

- Agility
- Flexibility
- Rapid Delivery
- Easy exploration
- Data acquisition is easier
- Data retrieval requires more effort

Data Warehouse

- Governance
- Reliability
- Standardization
- Security
- Data acquisition requires more effort
- Data retrieval is easier

Big Data Foundational Technologies

Big Data Technologies - "The Cloud"



Big Data Technologies - Public Cloud Services

1. Save capital expense and swap with variable expense
2. Benefits of the economies of scale of the cloud provider
3. Flexibility and Elasticity
4. Speed and agility
 - How long does it take you to get a server in your data centre?
5. Let a company who's 'day job' is running data centres run yours so your company can focus on their 'day job' which probably isn't running data centres

Big Data Technologies - Private + Public Cloud

- Hybrid is a mix of on premises and cloud
- Some applications and data cannot go to cloud
 - Regulatory compliance
 - Security requirements
- Many applications can move to the cloud
 - Make sure you have a fast and secure connection between on premises and cloud provider
- Hybrid is where most large corporates are right now

The Evolution of Databases

- Relational Databases & SQL
- The rise of NoSQL
- Popular NoSQL Databases

Relational Databases - SQL



A **relational database** is a digital database whose organization is based on the relational model of data, as proposed by E. F. Codd in 1970.

The **relational model** organizes data into one or more tables of columns and rows, with a unique key identifying each row.

Relationships are a logical connection between different tables, established on the basis of interaction among these tables.

Virtually all relational database systems use **SQL** as the language for querying and maintaining the database.

Relational Databases - SQL

- Relational Database Management System (RDBMS)
- Developed at IBM (early 1970s)
- First commercial version was by Relational Software (now Oracle) in 1979.
- SQL is an ISO standard, but most vendors add their own extensions.
- Popular implementations of SQL include
 - Oracle Database
 - MySQL – a free and open-source implementation now owned by Oracle
 - MS SQL Server
 - PostgreSQL

Relational Databases - Strengths

- Extremely well proven and widely used in the industry
 - E.g. Oracle, SQL Server, MySQL
- Quality of service guarantees
 - Highly efficient, e.g. via indexes, load balancing, etc.
 - Highly available, e.g. via replication, fail-over, etc.
 - Highly secure
 - Transactional

Relational Databases – Limitations

- Not good at storing unstructured or heterogeneous data
 - This kind of data doesn't fit nicely into the structured world of rectangular tables and fixed relationships
- Not ideal for ingressing big data at high velocity
 - It takes time to break the data down into rectangular chunks, so that it can be inserted into table(s) in an RDBMS
- Not good for rapidly evolving (agile) requirements
 - You can't keep changing the database schema all the time!
- Not ideal for scale-out architectures
- RDBMS aren't really designed for the cloud / commodity storage

The Rise of NoSQL

- The term "NoSQL" gained popularity around 2009
- NoSQL is a general term to represent non-relational database management systems
 - Encompasses a wide variety of database technologies
- NoSQL databases are designed to address the demands of building applications dealing with “Big Data”
 - Unstructured data
 - Handling big data
 - Data modelling agility
 - Scale-out architecture via auto-sharding, i.e. natively and automatically spread data across any number of servers

The Rise of NoSQL

● MongoDB
Topic

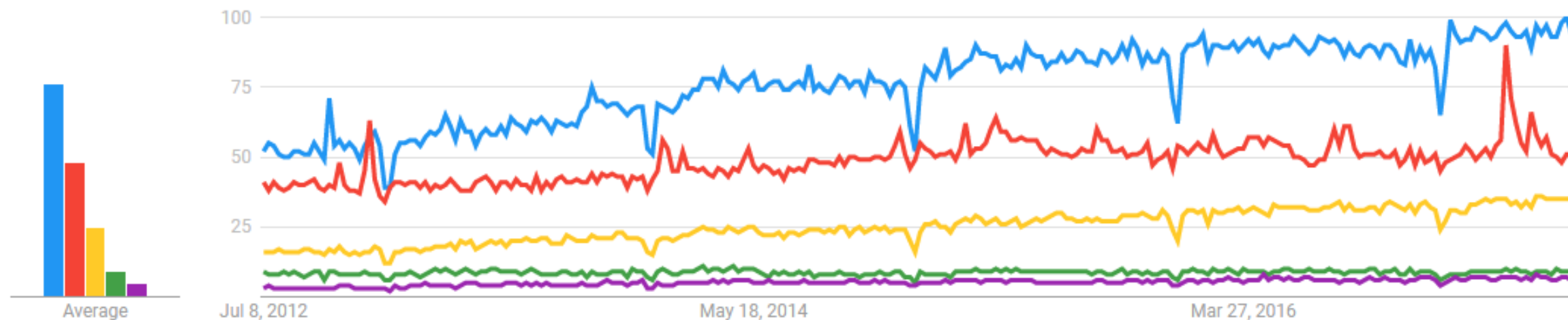
● Apache Cassandra
Topic

● Redis
Software

● Apache HBase
Software

● Neo4j
Search term

Interest over time ?



Data Models: Relational to Document

Relational

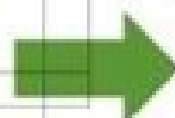
Person:

Pers_ID	Surname	First_Name	City
0	Miller	Paul	London
1	Ortega	Alvaro	Valencia
2	Huber	Urs	Zurich
3	Blanc	Gaston	Paris
4	Bertolini	Fabrizio	Rome

Car:

Car_ID	Model	Year	Value	Pers_ID
101	Bentley	1973	100000	0
102	Rolls Royce	1965	330000	0
103	Peugeot	1993	500	3
104	Ferrari	2005	150000	4
105	Renault	1998	2000	3
106	Renault	2001	7000	3
107	Smart	1999	2000	2

no relation



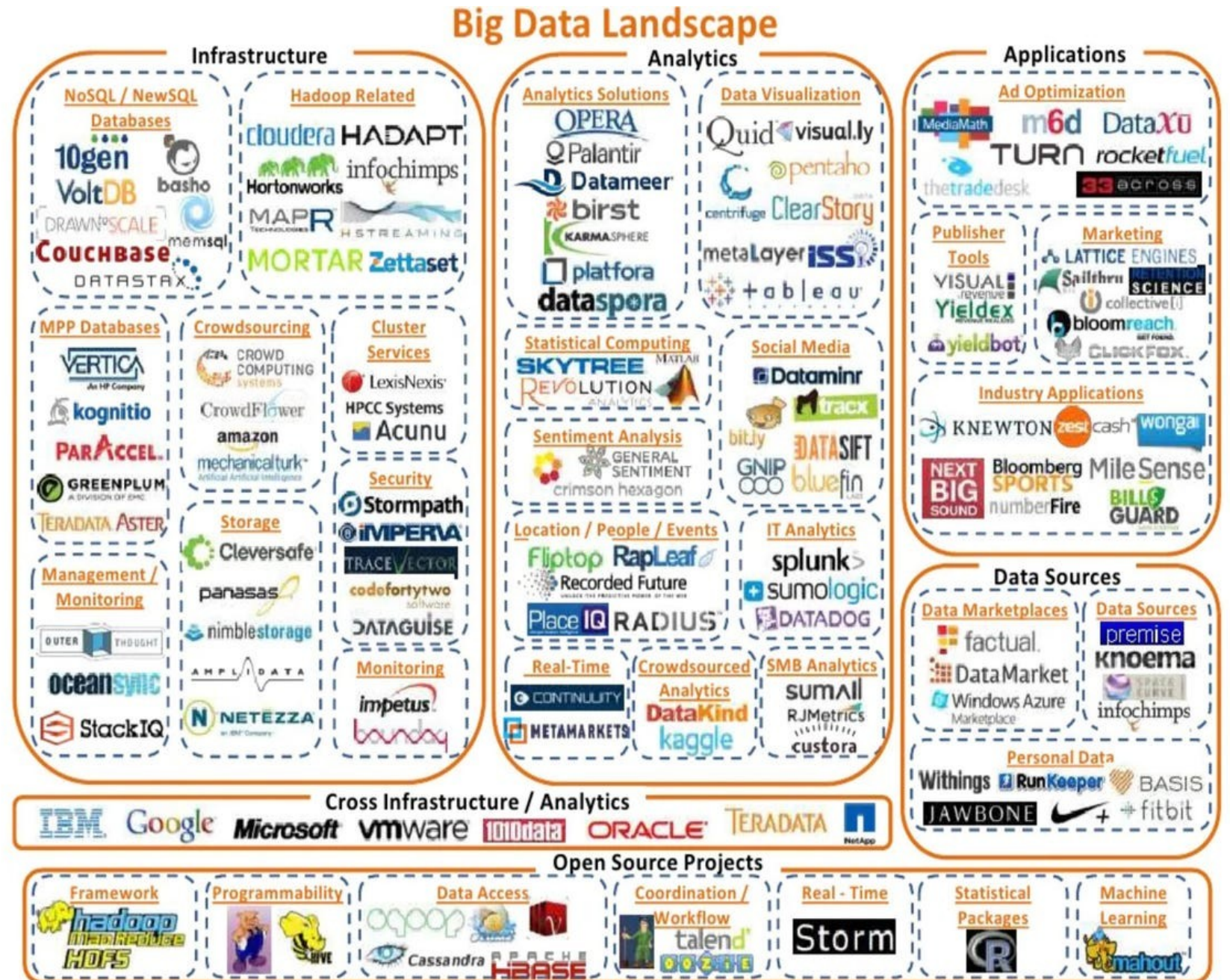
MongoDB Document

```
{  
  first_name: 'Paul',  
  surname: 'Miller',  
  city: 'London',  
  location: [45.123,47.232],  
  cars: [  
    { model: 'Bentley',  
      year: 1973,  
      value: 100000, ... },  
    { model: 'Rolls Royce',  
      year: 1965,  
      value: 330000, ... }  
  ]  
}
```

Big Data Technology Landscape

- Public Cloud
- Distributed Storage
- Distributed Analytics
- Unified Platforms
- Case Study: AWS S3

Technology for Big Data 2021



Distributed Storage

- Public cloud providers offer cheap data storage
 - Structured
 - Non-structured
- High scalability and availability
- Set up for analytics and machine learning

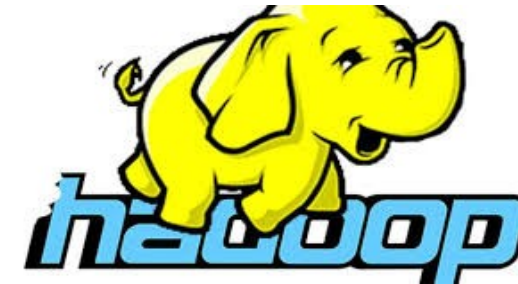
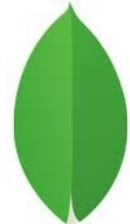


Azure Blob Storage



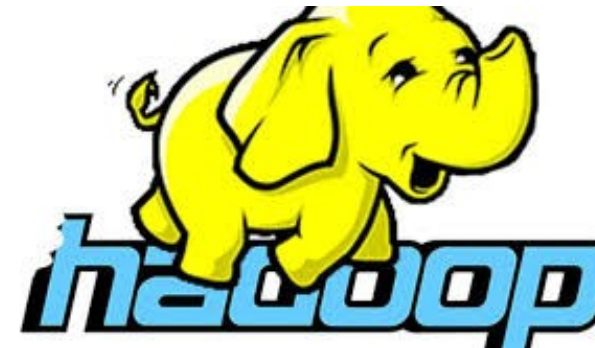
Google Cloud Storage

mongoDB®



Distributed Processing

- A number of technologies facilitate the collection and processing of large and/or rapidly growing data sets in batch to real-time
- High scalability and availability
- Set up for analytics and machine learning



Unified Platform Offerings



- More and more providers are offering a single platform which provides some or all of the following
 - data storage in the form of data lakes AND data warehouses
 - distributed analytics of the stored data
 - Transparently store data across public and private infrastructure

Case Study: AWS S3

- For Object Storage, AWS offers
 - S3
 - General purpose storage
 - Immediately accessible
 - Can be made available online
 - Glacier
 - Archive storage
 - Not immediately accessible
 - Cannot be made available online
 - The cheapest storage option of all



Case Study: AWS S3 Infrastructure

- There are 20 regions around the globe & 5 coming soon*

*as of February 2020

- Regions
- Coming Soon
- GovCloud



Case Study: AWS S3



- S3 provides object storage for files of up to 5TB in size
- The storage is unlimited
- High durability (99.99999999%)
 - 11 nines of durability achieved by
 - Files saved on multiple drives across multiple data centres
 - *“If you store 10,000 objects with us, on average we may lose one of them every 10 million years or so. This storage is designed in such a way that we can sustain the concurrent loss of data in two separate storage facilities.”*

Case Study: AWS S3



- When using S3 you create **buckets** into which you place **objects**
- These objects have a **key** which is the object identifier
- Buckets can be made available over the Internet
 - Static Web sites can be hosted from buckets
 - CSS / JS / Images can be hosted from buckets

Case Study: AWS S3



- Data can be uploaded
 - Over the Internet
 - Over Direct Connect
 - Uploaded using SnowBall
 - Uploaded using SnowMobile



- Often S3 storage is used as the core of a data lake

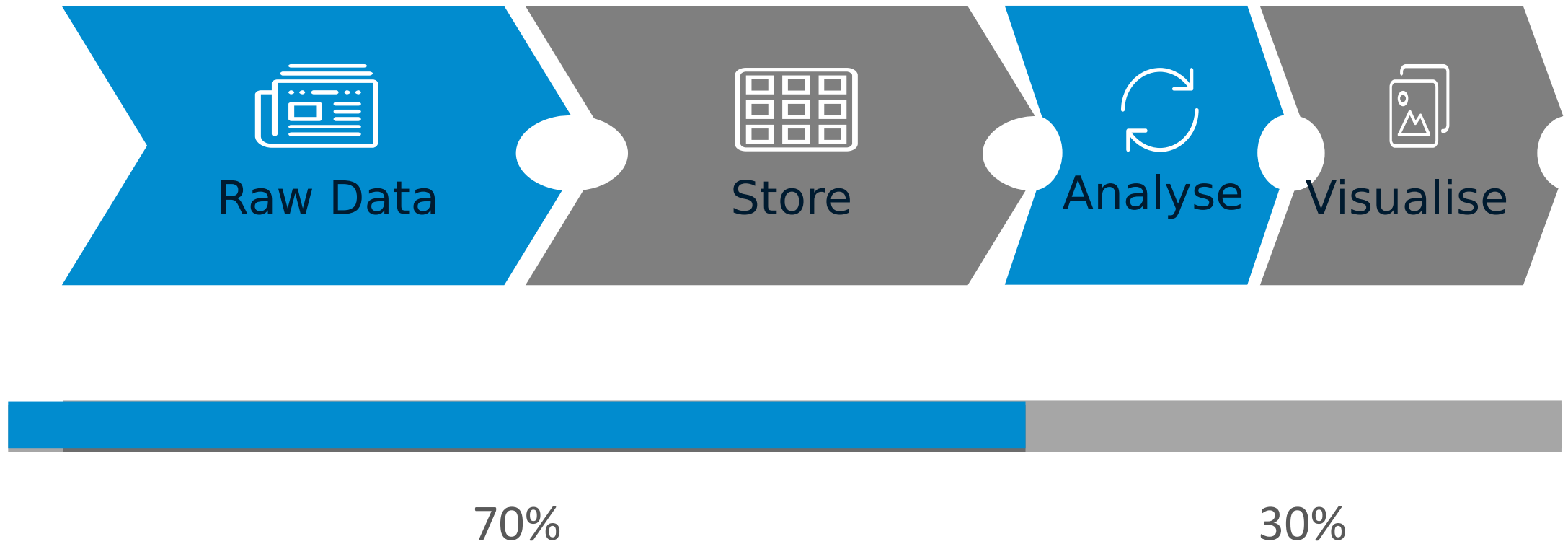
Case Study: AWS S3 Lab Exercises



- Sign up / Sign in at <https://amazon.qwiklabs.com/>
 - Introduction to the Relational Database Service
 - Introduction to S3

Data Science Projects

The Data Science Process - Effort



ETL – Extract, Transform, Load

- (**ETL**) is the general procedure of copying data from one or more sources into a destination system which represents the data differently from the source(s) or in a different context than the source(s)
- Traditionally, ETL has been used to move data between elements in a data pipeline
 - Online Transaction Processing Database => Data Lake => Data Warehouse
- More recently, unified data platforms provide a "one-stop-shop" thus reducing the need for ETL

Data in the real world is dirty

- **Incomplete:**
 - lacking attribute values
 - lacking certain attributes of interest
- **Noisy:**
 - containing errors or outliers (spelling, phonetic and typing errors, word transpositions, multiple values in a single free-form field)
- **Inconsistent:**
 - containing discrepancies in codes or names (synonyms and nicknames)
 - prefix and suffix variations
 - abbreviations, truncation and initials
- **Lack of Currency**
 - Out of date
 - No longer relevant

Why is Data Dirty?

- Incomplete data comes from:
 - non available data value when collected
 - different criteria between the time when the data was collected and when it is analyzed.
 - human/hardware/software problems
- Noisy data comes from:
 - data collection: faulty instruments
 - data entry: human or computer errors
 - data transmission
- Inconsistent (and redundant) data comes from:
 - Different data sources, so non uniform naming conventions/data codes
 - Functional dependency and/or referential integrity violation

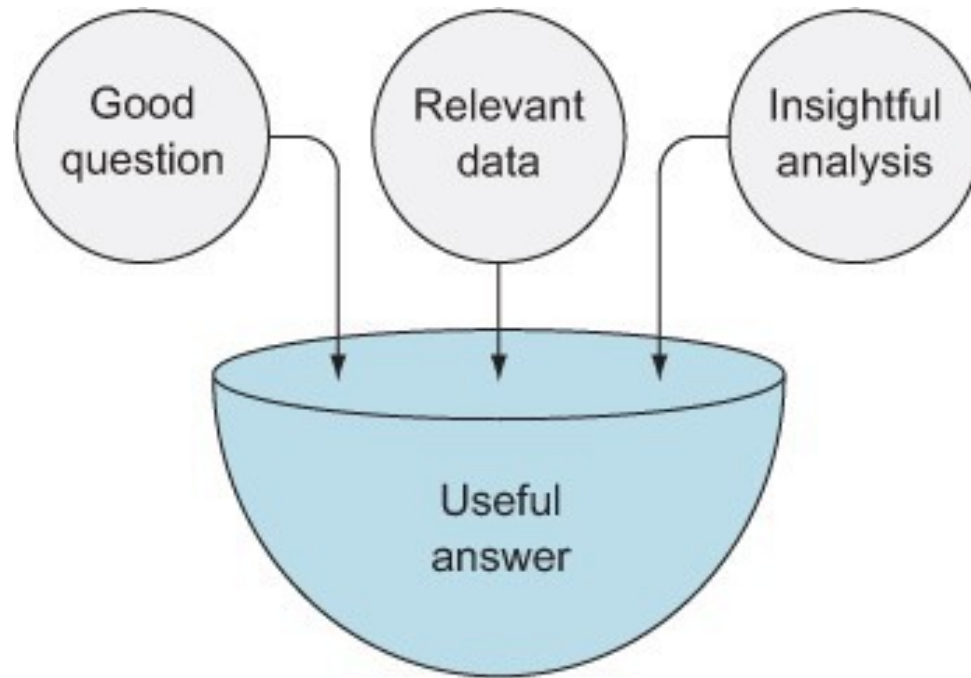
What do we need to ask about our data?

- Can we interpret the data?
 - What do the fields mean?
- Are there data glitches?
 - Typos, multiple formats, missing / default values
- Do you need metadata and/or domain expertise
 - Is the revenue field in cents, euros, or 1000s euros?

Cleaning Tools

- Most data analytics engines will include specific features for cleaning data
- "No Code" Data Scrubbing Tools:
 - help to clean and correct lists and databases by identifying incomplete, incorrect, inaccurate, irrelevant, etc. parts of the data and then replacing, modifying, or deleting this dirty data
 - E.g. XPlenty, Tibco Clarity
- Scripting Languages:
 - Often scripts in Python and R are used to do bulk clean-up.
 - There are libraries available to help e.g. treat all empty cells in the same way (e.g. fill with average, media, or specific values, or be removed entirely)
 - There are libraries to enable processing of distributed data (e.g. Spark)

Data Analysis



Getting Started on Projects

- Start small and iterate
- Be question or hypothesis led
- Solve real problems that address a business challenge and are actionable
- Involve all stakeholders
- Build partnerships between functions (IT, finance, business)
- Consider collecting data from new sources such as short surveys or other qualitative information
- Tell a story from a data

Don't Confuse Reporting with Analysis

Reporting



Analysis

- Asking questions and finding answers in the data
- Forming hypothesis, testing and measuring results in the data.
- Understanding the meaning behind the numbers and lines and taking action based on that new understanding

Failure Rate of Data Projects

- July 2019: VentureBeat AI reports **87% of data science projects never make it into production**
- Jan 2019: New Vantage survey reports 77% of businesses report that "business adoption" of big data and AI initiatives continues to represent a big challenge for business
- Jan 2019: Gartner says 80% of analytics insights will not deliver business outcomes through 2022

Reasons Projects Fail

- The question is not relevant to the business
- The timeline is too ambitious
 - Look for short term wins and build the analysis as you go
- Poor quality data
 - Fields are mislabeled or abbreviated in different ways across the company
- Failure to address legal, compliance, privacy and ownership issues
- Not possible to translate insights into action
- The data is not presented in a way that is meaningful to the consumer of the insights

Questions to Ask Prior to Starting A Data Project

The Problem

1. What is my business question?
2. What is the current solution to the problem? Will the proposed approach be i) better and ii) worth the cost of investment ?
3. Who are key stakeholders and what are their most important questions?
4. Who is the audience?
5. Who can access the information?
6. How will the results be used?
7. What reports will be produced? Detail the data and charts on each report.
8. What is the simplest and most effective solution to the problem that can be created quickly? i) one-off answer to support a strategic decision, ii) standalone light-weight app for stakeholders to use iii) a real-time data product that integrates into other systems?

Questions to Ask Prior to Starting A Data Project

The Data

1. What data sources are available to work with?
2. Will this data answer my business question?
3. What is size of each data set and how much data will you need to use from each one?
4. For each individual source – i) Is it complete? ii) Accurate? iii) Up to date?
5. Do you have the required permissions or credentials to access the data necessary for analysis?
6. What transformation needs to be done on the data?
7. What is the frequency of updates required for the data?
8. Who will maintain the data?
9. Have you addressed all legal and compliance issues related to this data?

Questions to Ask Prior to Starting A Data Project

The Project

1. How will success be measured?
2. What is the ROI?

Data Policies

- Create clear, consistent policies that are readily available and easy to understand
 - Transparency
 - Consent
 - Privacy
 - Anonymization
 - Usage and aggregation
- Encrypt and secure all data
- Aggregate data and report on aggregated results

Thankyou
