PeopleCert Data science

PeopleCert Data Science

Foundation

Three-Day Course

Study Guide

Copyright Details

e-mail: info@peoplecert.org, www.peoplecert.org

- Document Version: PC-DS_FND_SG 1.0 | February 2022

PeopleCert
All talents, certified.

PeopleCert: A Global Leader in Certification

- ✓ **Web** & **Paper based** exams in **25** languages
- ✓ **2,500 Accredited Training Organizations** worldwide
- ✓ Delivering exams across **200 countries every year**
- ✓ Comprehensive Portfolio of **500+ Exams** and Growing

AXELOS GLOBAL BEST PRACTICE · PeopleCert DevOps. · PeopleCert Scrum. · PeopleCert DEVSkills. · PeopleCert Lean IASSC · PeopleCert LeanIT. · ISACA · IASSC · Language Cert · ECDL

PeopleCert
All talents, certified.

How to Use This Document

This document is your **PeopleCert Data Science Foundation Study Guide** to help you prepare for the **PeopleCert Data Science Foundation examination (Essentials/Beginner level)**.

It is meant to provide you with a clear outline of everything covered in the course presentation by your instructor that will be on the PeopleCert Data Science Foundation exam.

Your exams will be closed book. You will be given 60 minutes to complete it. It contains 40 multiple choice questions and to pass the exam you must achieve a grade of 70% or higher, or a minimum of 28/40 correct responses. For further details on your exam, including more information on question types and learning objectives, please refer to your course syllabus.

As you follow along, you may see that some material here is not replicated in the trainer presentation. This study guide includes questions, activities, knowledge checks, or other material in the presentation that are facilitated verbally by the instructor. It also does not contain content that is not examinable, but instead is designed to reinforce learning or add value to your course experience. It also provides valuable links and references, throughout the slides, which you can explore further to enhance your learning and understanding of the material provided in the study guide.

PeopleCert
All talents, certified.

Syllabus

Category	Topic	Skill Set
1.0 Introduction to Data Science	1.1 Overview & Definitions	1.1.1 General Terms and Definitions
		1.1.2 Big Data
	1.2 Key Concepts	1.2.1 Data Analytics
		1.2.2 Data Science
		1.2.3 Machine Learning
		1.2.4 Artificial Intelligence (AI)
2.0 Programming Skills (with R/Python)	2.1 Introduction to Programming	2.1.1 Key Concepts
3.0 Data Management	3.1 Relational Databases (RDBMS)	3.1.1 Key Concepts
	3.2 New Data Management Methods	3.2.1 NoSQL
	3.3 Business Intelligence	3.3.1 Key Concepts
4.0 Probability & Statistics	4.1 Introduction to Statistics	4.1.1 Key Concepts
	4.2 Introduction to Probability Theory	4.2.1 Key Concepts of Probability Theory
5.0 Machine Learning (ML) and Artificial Intelligence (AI)	5.1 Machine Learning (ML)	5.1.1 Introduction to ML
6.0 Visualization	6.1 Introduction to Visualization	6.1.1 Key Concepts of Visualization
7.0 Business Skills	7.1 Data Governance	7.1.1 Key Concepts of Data Governance
	7.2 Ethics, Data Privacy and Protection	7.2.1 Data Privacy & GDPR

PeopleCert Data Science Foundation

Introduction

Data Science Foundation | Welcome

Welcome to a three-day course that provides basic knowledge and understanding of Data Science concepts for candidates who have no technical background and/or related experience, primarily professionals of any field and in general anyone who wants to obtain awareness of the basic principles of Data Science.

Image Source: https://towardsdatascience.com/why-data-science-succeeds-or-fails-c24edd2d2f9

Lesson Plan | Agenda

- **Day #1**
  - Introduction to Data Science
  - Why is Data Science and Data important?
  - Common Tools used in Data Science
  - Data Science and Programming
  - Recap and Q&A
- **Day #2**
  - Data Management
  - Data Governance
  - Business Intelligence & Visualization of Data
- **Day #3**
  - Statistics
  - Machine Learning (ML)
  - Artificial Intelligence (AI)
  - Recap and Q&A

PeopleCert Data Science Foundation

Day #1

Introduction to Data Science

Objectives:	Syllabus Topics:
▪ Introduction to Data Science	▪ 1 Introduction to Data Science
▪ Why is Data Science and Data important	▪ 2 Programming Skills (with R/Python)
▪ Common Tools used in Data Science	
▪ Data Science and Programming	

PeopleCert Data Science Foundation

Introduction to Data Science and Why is Data Science and Data Important

Why is Data and Data Science Necessary?

- Consider the terms “data” and “data science”. Do these seem scary to you?
- Then consider a world without data…how would that implicate your business life?
- Data science provides businesses with the ability to process and interpret data in order to make informed decisions around several aspects of the business including growth, optimization, customer satisfaction, internal evaluation of financial goals and performance.
- Data Science combines statistics, mathematics, computation (and computing) and analyses large amounts of complex and raw organizational data and provides meaningful information based on that data to the company. It is a combination of many fields such as statistics, mathematics, and computation to interpret and present data for effective decision-making by business leaders.

What is Data Science

Definition:

'*Data science combines the scientific method, math and statistics, specialized programming, advanced analytics, AI, and even storytelling to uncover and explain the business insights buried in data.* '

Data science encompasses:

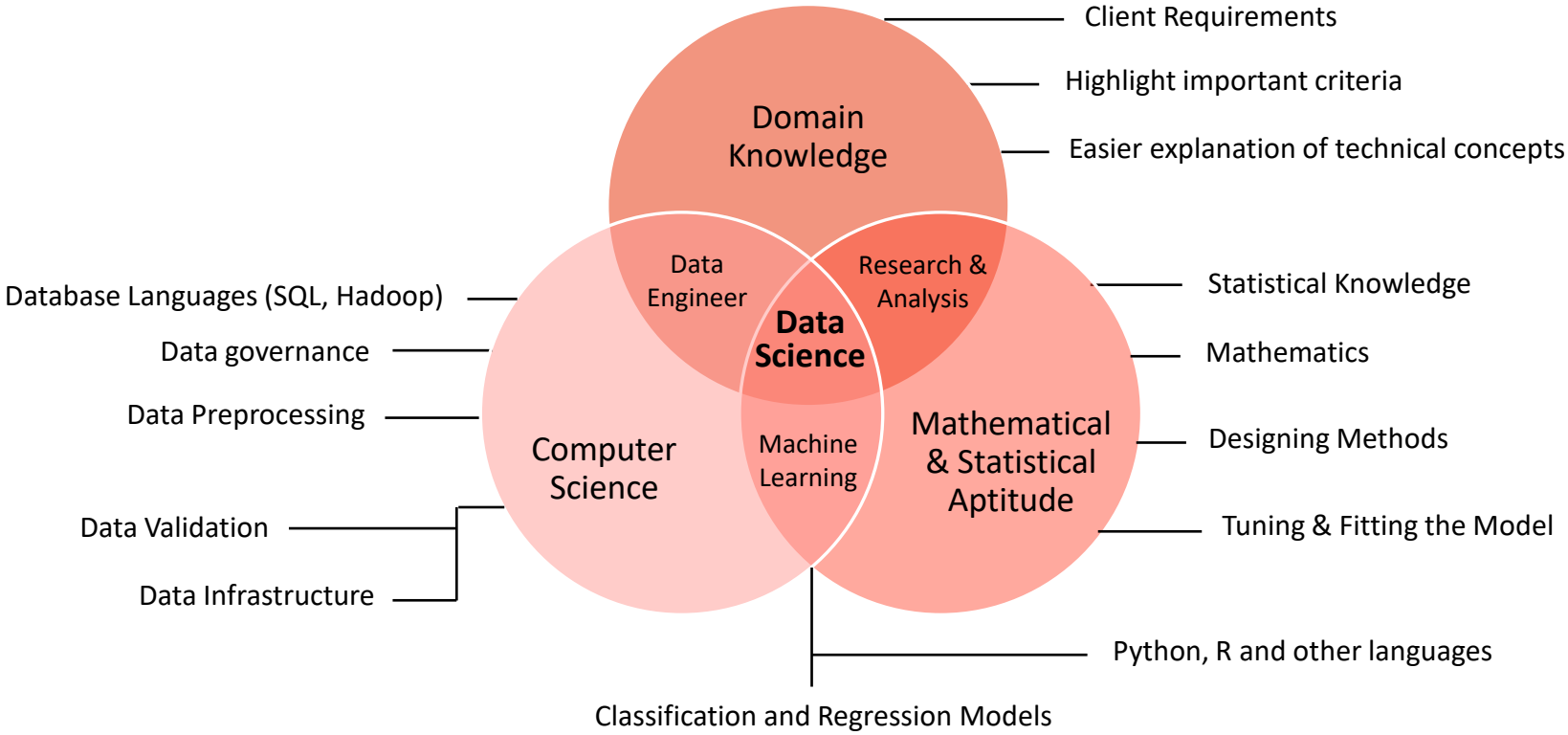
- **preparing data** for analysis and processing
- performing **advanced data analysis** using complex algorithms, analytics, Artificial Intelligence and Machine Learning to find patterns
- **transforming** these patterns into **predictions** that support business decision making
- **validating the results** through scientifically designed tests and experiments
- **presenting the results** in visualizations that enable stakeholders to draw informed conclusions

Data Scientists are both rare and in high demand. As different organizations provide different job role descriptions to data science related jobs, it is important to distinguish between jobs that require pure modelling use cases and jobs that also need code development and deployment skillsets.

Source: https://www.ibm.com/cloud/learn/data-science-introduction

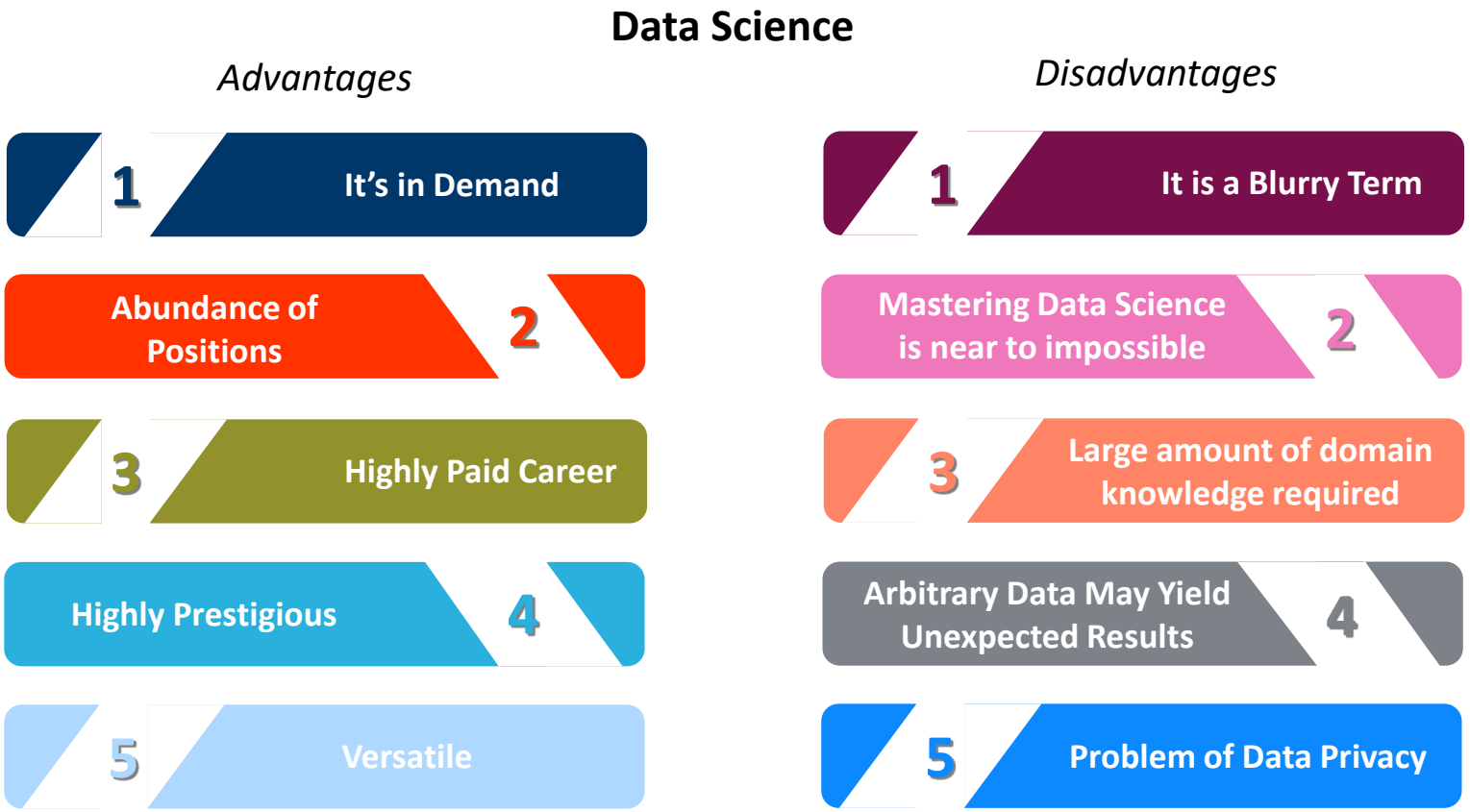
What is Data Science?

"Data Science is about extraction, preparation, analysis, visualization, and maintenance of information. It is a cross-disciplinary field which uses scientific methods and processes to draw insights from data."



Source: https://data-flair.training/blogs/what-is-data-science/

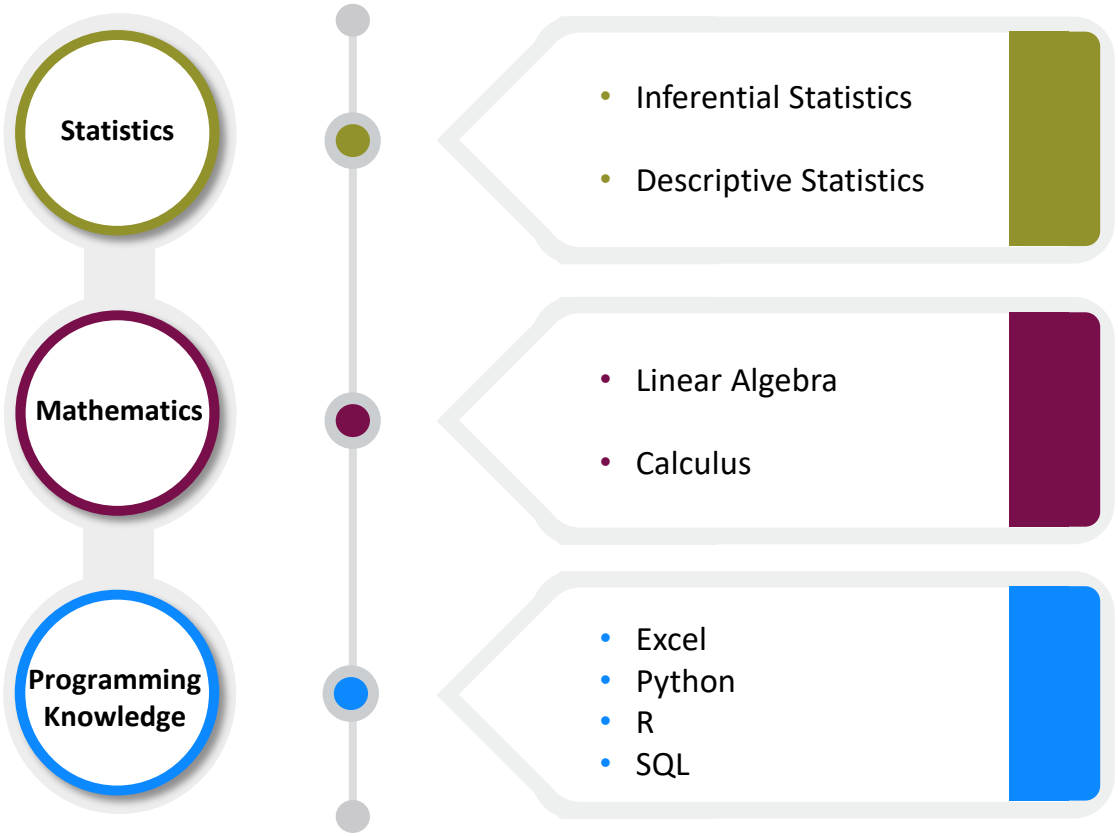
Pros and Cons of Data Science



Source: https://data-flair.training/blogs/pros-and-cons-of-data-science/

Data Science Skills

Data Science Prerequisites



Source: https://data-flair.training/blogs/data-science-prerequisites/

Why is Data Science Important

- Data has become the fuel of industries as companies require data to function, grow and improve their businesses
- Analysing a large amount of data to derive meaningful insights, with the help of Data Scientists assist companies in making valuable (and rationalized) decisions
- Insights derived from data analysis assist companies to analyse themselves and analyse their performance in the market
- Real world problems can be solved with Data Science
- The number of roles for Data Scientists has grown by 650% since 2012.
- About 11.5 Million jobs will be created by 2026 according to the U.S. Bureau of Labor Statistics.
- The job of Data Scientist ranks among top emerging jobs on Linkedin
- Data Science has applications in several fields, that will be detailed below
- Data Science is a very robust field that best fits people who have a knack for experimentation and problem-solving

Source: https://data-flair.training/blogs/what-is-data-science/

How Could Data Science Help you Professionally?

- Empowers you to make better decisions
- Helps you focus on the important core issues
- Steers you in the right direction as it can predict and track upcoming trends
- Helps you identify new opportunities
- Provides evidence to your decisions

Data science adds business value to your organization and makes you, as a professional, make better, informed and evidence-based decisions!

Source: https://www.upgrad.com/blog/why-data-science-important/

The Stages in the Data Science Lifecycle

The lifecycle "wheel" isn't set in stone.

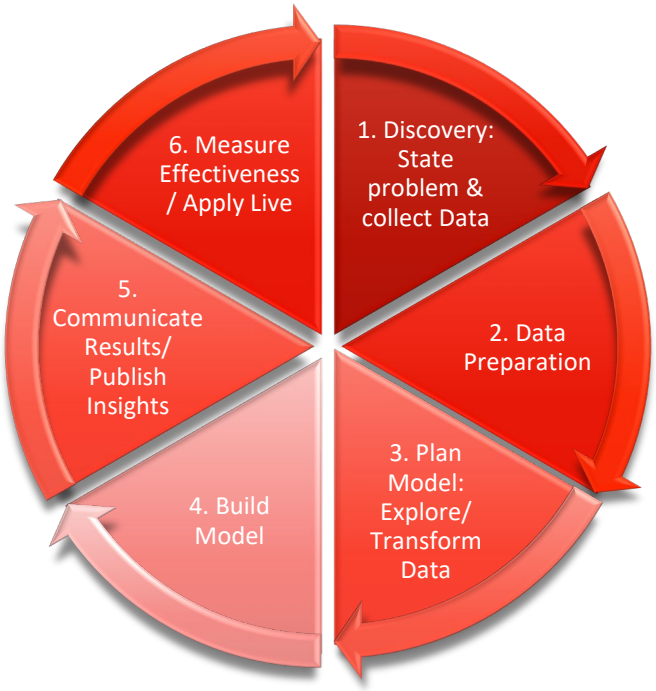
While it's common to move through the phases in order, it's possible to move in either direction (i.e.., forward, backward) at any stage in the cycle.

Work can also happen in several phases at the same time, or you can skip over entire phases. In addition, if new information is uncovered, work can return to an earlier phase to start the cycle over again.

Notes:

- The term "data lifecycle" is also up for debate, as data doesn't really evolve and grow like a seed or egg would. Some authors add different stages.
- For example, purging. That addition might not be accurate, as it's not common for data to be deleted out of existence; It's much more likely to be stored or archived (the equivalent of suspended animation).
- Different people may call different parts of the wheel something slightly different.
- For example, "data prep" might be called "data capture".

This simplified model gives you a starting point with which to build a data lifecycle that works for your organization.



Source: https://www.datasciencecentral.com/profiles/blogs/the-lifecycle-of-data

What is Big Data

Definition:

'*Big data refers to the 21st-century phenomenon of exponential growth of business data, and the challenges that come with it, including holistic collection, storage, management, and analysis of all the data that a business owns or uses*.'

What kind of data does Big Data imply?

The term implies data of an indeterminate—and steadily increasing—size as well as from an indeterminate number of sources, including data generated by employees, customers, partners, machines, logs, databases, security cameras, mobile devices, social media, and more.

What advantages does Big Data offer?

- **Scalability:** computation and storage scale linearly
- **Flexibility**: no fixed data formats needed
- **Limitless:** no predefined limits on storage and computation
- **Coping with uncertainty:** changes in data formats are supported without issues
- **Characteristics of big data: Volume, Velocity, Variety, Veracity, Value** *(are presented in greater detail in the next slides)*

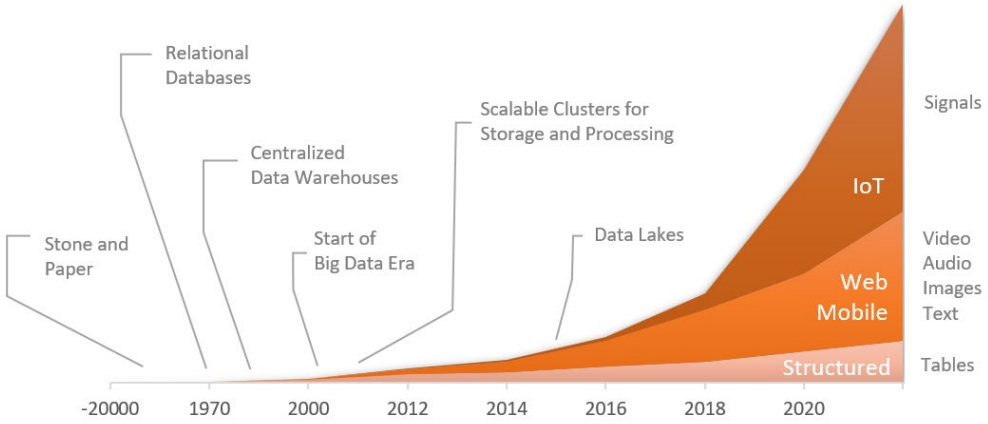
Why is big data a challenge?

- The **amount** of data being generated **keeps increasing**, so the challenge to contain big data is a moving target.
- A **competitive** problem: As more companies invest in, and have success with, their big data, those companies that do not keep step will be at great disadvantage.
- The **technologies** and commercial products to contain and control big data **are evolving rapidly**, so IT organizations must stay alert to new innovations and opportunities.

Source: https://www.informatica.com/services-and-training/glossary-of-terms/big-data-definition.html
https://www.researchgate.net/publication/333171951

Explosion of Data

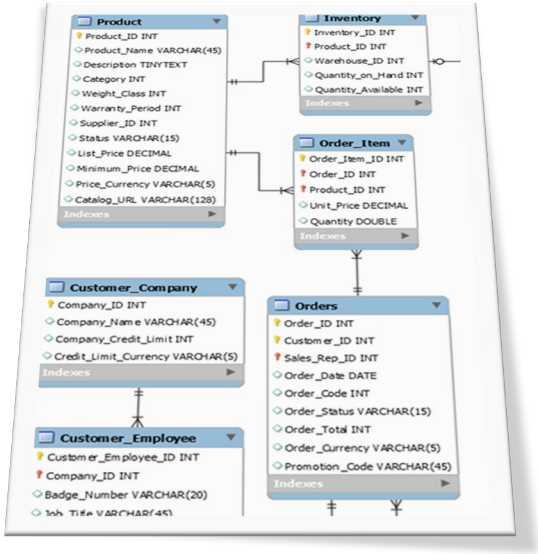
- **History**: Societies were more successful if they were able to store and use data
  - Couldn't handle growth in physical storage
  - **Example:** tribespeople (18000 BCE) mark notches into sticks or bones, to keep track of trading activity
  - **Example:** The abacus (C 2400 BCE) The first dedicated device constructed specifically for performing calculations – comes into use in Babylon
  - **Example:** The Library of Alexandria (300 BC – 48 AD) The largest collection of data in the ancient world, housing up to perhaps half a million scrolls and covering everything we had learned so far
- **1970 – 2000**: Data grows exponentially, new possibilities to store and search data on mainframes and later in Relational Databases
  - **Example:** OLTP, OLAP, RDBMS
- **2000 – Now**: Big Data with scalable solutions for storage, processing and streaming



Source: https://insidebigdata.com/2017/02/16/the-exponential-growth-of-data/

Before Big Data

- Traditionally, **Data** had to provide **data format consistency**.
- A **Database** holds the data in a controllable structure. It is a **centralized** place for a fixed schema of columns, tables and relations.
- The database is used (with SQL) for a transactional **Application** (OLTP) for fast **operational processing** or a traditional **Data Warehouse** (OLAP) for **analytical queries**.
  - Online transaction processing (**OLTP**) captures, stores, and processes data from transactions in real time in (transactional) operational applications.
  - Online analytical processing (**OLAP**) uses complex queries to analyse aggregated historical data from OLTP systems. A datawarehouse is an example of OLAP environment for analytical queries.
  - The data from one or more OLTP databases is ingested into OLAP systems through a process called extract, transform, load (**ETL**).
- But this is also rigid, inflexible and has limited scale. Structure and certainty comes at a price, and that **price is scalability** (especially in some more antiquate technologies since modern day cloud technologies are much better and at a better price).



Source: https://www.stitchdata.com/resources/oltp-vs-olap/

The Rise of Big Data

From **2000 to 2012** the landscape changed towards Big Data, triggered by exponentially growing data.

Web, Social & Mobile use:

- Google started indexing the web
- Yahoo, Amazon and eBay started analyzing click rates
- Mobile GPS data applications
- Facebook connected user profiles into a graph structure
- Netflix network traffic usage and recommendation engine

This required not-yet invented approaches for platforms and tools to order and analyse these **new data scenarios**.

Early 2000s

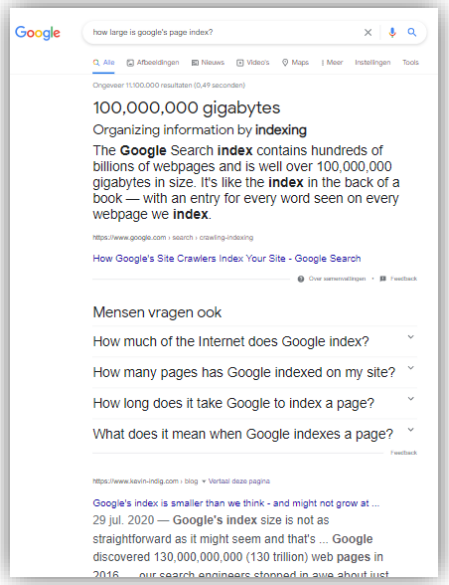
- New Data scenarios occurred due to increased web/mobile use. Large companies had to find new approaches and storage solutions to deal with these new data types in order to analyse them effectively
- Techniques that arrived to analyse data in a decentralized way: MapReduce, Hadoop and BigQuery

2012

- Hadoop 1.0 became a mature solution

More recently

- IoT scenarios contributed to new BigData scenarios with sensor-based internet-enabled devices
- Self-driving cars, like from Tesla, analyse real-time video-capturing of the road/traffic



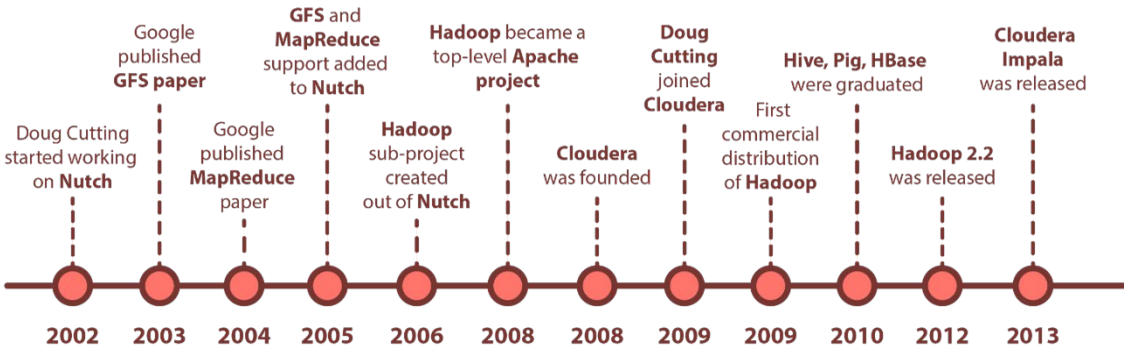
Maturity of Big Data

- Between **2012 to 2018** the key players in **the market changed** almost every year.
- **Since 2018**, the market has been stabilized and the large Clouds have reached the point where they can offer a stable and **mature landscape** that makes traditional value of data (SQL, DWH) and innovation and flexibility (Big Data, AI / ML) available to everyone.
- **More recently full-scale IoT** contributed to new Big Data scenarios with sensor-based internet-enabled devices.

Big Data Landscape 'The Beginning' (1/2)

Early 2000's the Big Data landscape started to exists and grow, this was triggered by:

1. Increased web usage created a **need to search** through webpages
2. This triggered in 2003 the creation of the Apache Notch web mining platform for **crawl-index-search**
3. The platform added Google's MapReduce for data processing on **large scale clusters** and the **Distributed File System** for storage
4. MapReduce and NDFS were in 2006 the start of the **Hadoop ecosystem** (with Hive, Hbase, Pig, etc..) with a stable release in 2012



Big Data Landscape 'The Beginning' (2/2)

A brief history

1. Apache Notch (2003) Apache Web Mining Platform: Open Source web search engine for crawling+indexing+searching through webpages, initiated by Yahoo!
2. MapReduce (Google 2004) Simplified data processing on large scale clusters
3. Distributed File System (GFS, Google 2003 / NDFS, Notch 2005) to store data in a decentralized way
4. In 2005 NDFS and MapReduce were added as the storage and processing part for Notch
5. In 2005 Apache Notch became a subproject of Lucene (search engine)
6. In 2006 NDFS and MapReduce were taken into a separate Apache subproject: Hadoop
7. In 2009 Cloudera started to offer a commercial distribution of Hadoop
8. Hadoop ecosystem:
    1. HBase is a key value store (mostly)
    2. Hive is a system to execute SQL-like queries on a Hadoop system
    3. Pig is a special query language to access big data
9. In 2012 Hadoop V2 was released, this was the first workable version. It contained YARN, which splits up the functionalities of resource management (which does job tracking and resource allocation to applications) and job scheduling/monitoring (which monitors progress of the execution). The idea is to have a global ResourceManager (RM) and per-application ApplicationMaster (AM).