
JAMES COOK UNIVERSITY
AUSTRALIA
Week 1
MA5831 – Advanced Data Management and Analysis using SAS
Dr Mostafa Shaikh
mostafa.shaikh@jcu.edu.au
online.jcu.edu.au
Cairns
Singapore
Townsville

Agenda

- Course outline
- Week 1: Data for processing and analysis
  - Big Data – future landscape
  - Data Governance and Data Analysis Framework
  - Data processing in enterprises
  - Data processing techniques
  - Data storage technologies

- Technical setup
  - SAS Data Academy and VM

Course outline

JAMES COOK UNIVERSITY AUSTRALIA

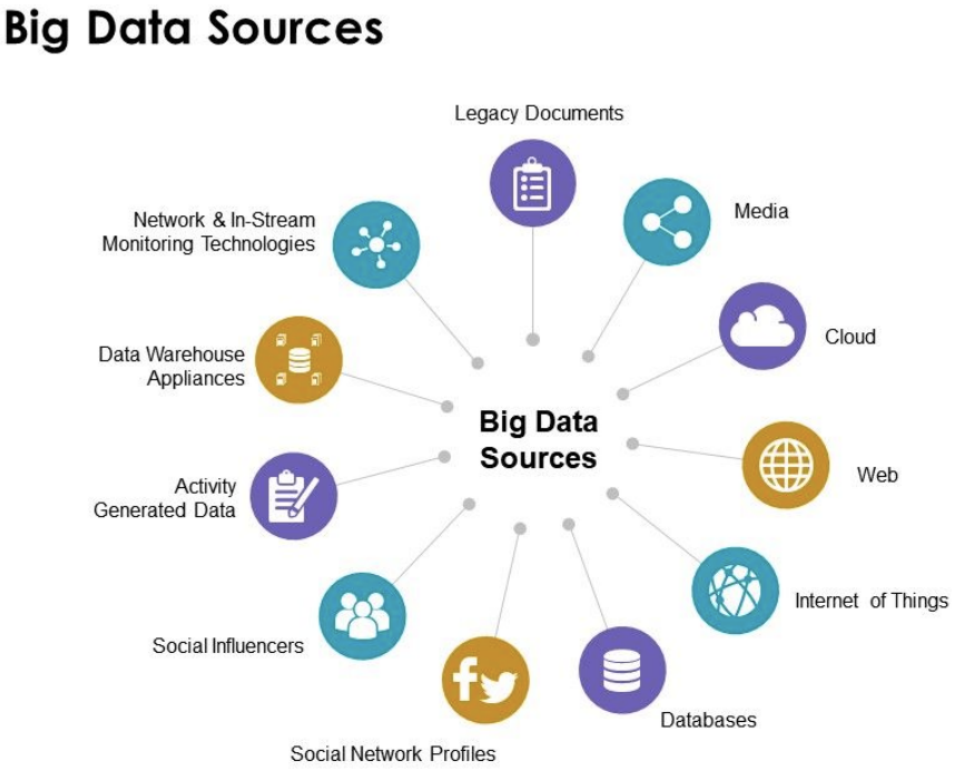
- Content
  - W1 – W3: Data Engineering (data processing, cleansing, parsing: Plan, Act, Monitor)
    - SAS Data Management Studio (2.6)
  - W3 – W6: Big Data ecosystem
    - SAS Cloudera
- Assignment:
  - A1: Data Processing Quiz (10Q, 10%)
  - A2: Data quality profiling and standardising (case study, 20%)
  - A3: Data processing trends – Cloud data warehouse (lit review, 40%)
  - A4: Managing data with Hive and Pig in Hadoop (case study, 20%)
  - A5: SAS Academy of Data Science course chapter completion (online material review, 10%)

Foundation of standard data

PC era: Database [DB] and Database Management System [DBMS] using Relational Database Models [RDBM]

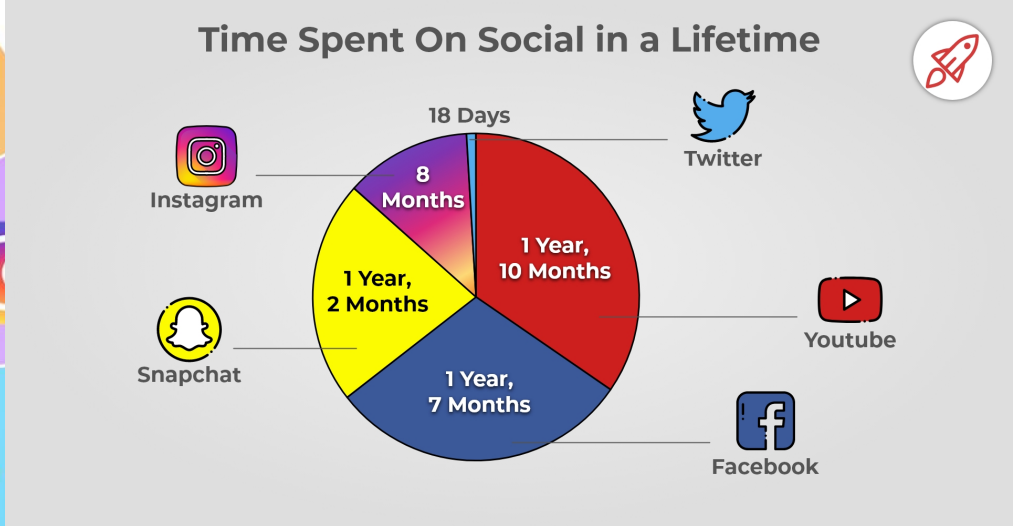
Internet era: better linkage of data over the web (e.g., XML, RDF) along with improved protocols for data distribution (e.g., blockchain)

Data era: new level of data complexity (variety), velocity and volume which is known as ‘big data’



Why we need to innovate?

in 2020, some 3.8 million people use social media, which is more than half the world's population.



“By 2025, nearly 30 percent of data generated will be real-time”

Why we need to innovate?

JAMES COOK UNIVERSITY AUSTRALIA

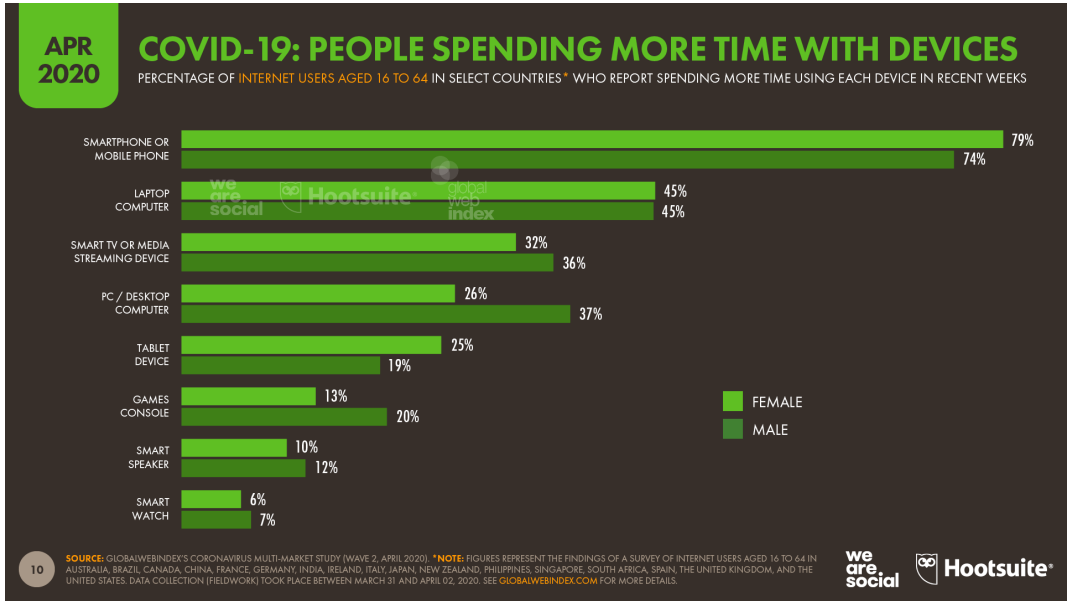
IDC's top 10 worldwide IT industry predictions

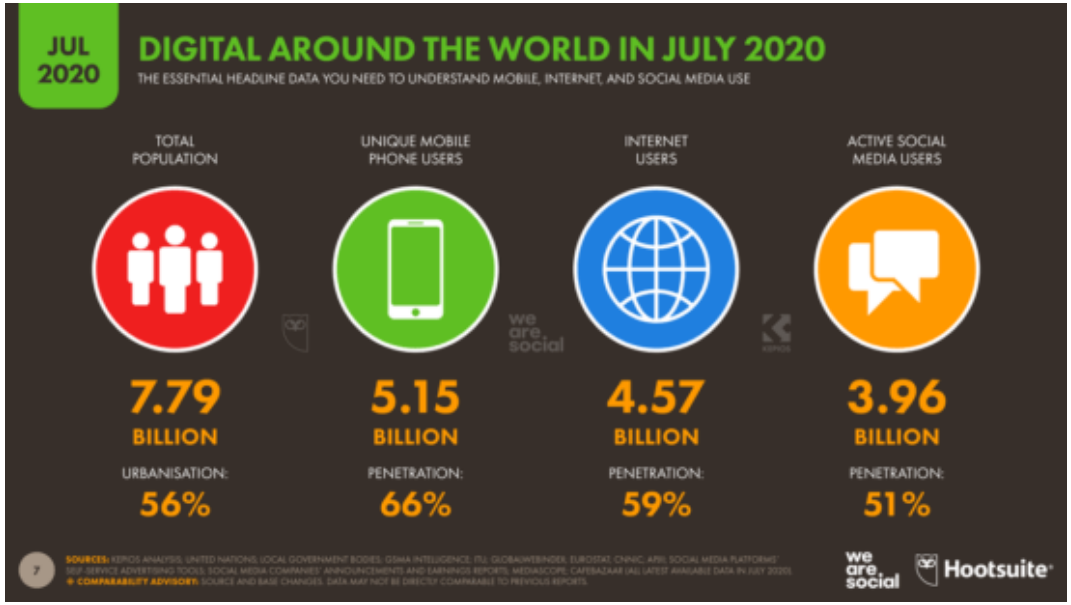
1. By 2022, over **60% of global GDP** will be **digitized** with growth in every industry driven by digitally-enhanced offerings, operations, and relationships.
2. By 2023, 75% of all IT spending will be on 3rd Platform technologies, as over 90% of all enterprises build "**digital native**" IT environments to thrive in the digital economy.
3. By 2022, over 40% of organizations' cloud deployments will include **edge computing**, and 25% of endpoint devices and systems will execute **AI algorithms**.
4. By 2022, 90% of all apps will feature **microservices architectures** that improve the ability to design, debug, update, and leverage third-party code; 35% of all production apps will be **cloud-native**.
5. By 2024, a new class of professional developers **producing code without custom scripting**, will expand the developer population by 30%, accelerating digital transformation.
6. From 2018 to 2023, with new tools/platforms, more developers, agile methods, and lots of **code reuse**, **500 million new logical apps** will be created, equal to the number built over the past 40 years.
7. By 2022, 25% of public cloud computing will be based on non-x86 processors (including quantum); by 2022, organizations will spend more on **vertical SaaS apps** than horizontal apps.
8. By 2024, **AI-enabled user interfaces and process automation** will replace one third of today's screen-based apps. By 2022, 30% of enterprises will use conversational speech tech for customer engagement.
9. By 2022, 50% of servers will encrypt data at rest and in motion; over 50% of security alerts will be handled by AI-powered automation; and **150 million people will have blockchain-based digital identities**.
10. By 2022, the top four cloud "megaplatforms" will host 80% of IaaS/PaaS deployments, by 2024, 90% of G1000 organizations will mitigate lock-in through **multi- and hybrid cloud technologies and tools**.

https://www.techrepublic.com/article/10-predictions-for-how-the-tech-industry-will-change-in-2019-and-beyond/

Global trends driving big data

1. Mobile internet
2. Mass social media use
3. Global data growth
4. Open data movement
5. Low cost distributed computing
6. Sensor data

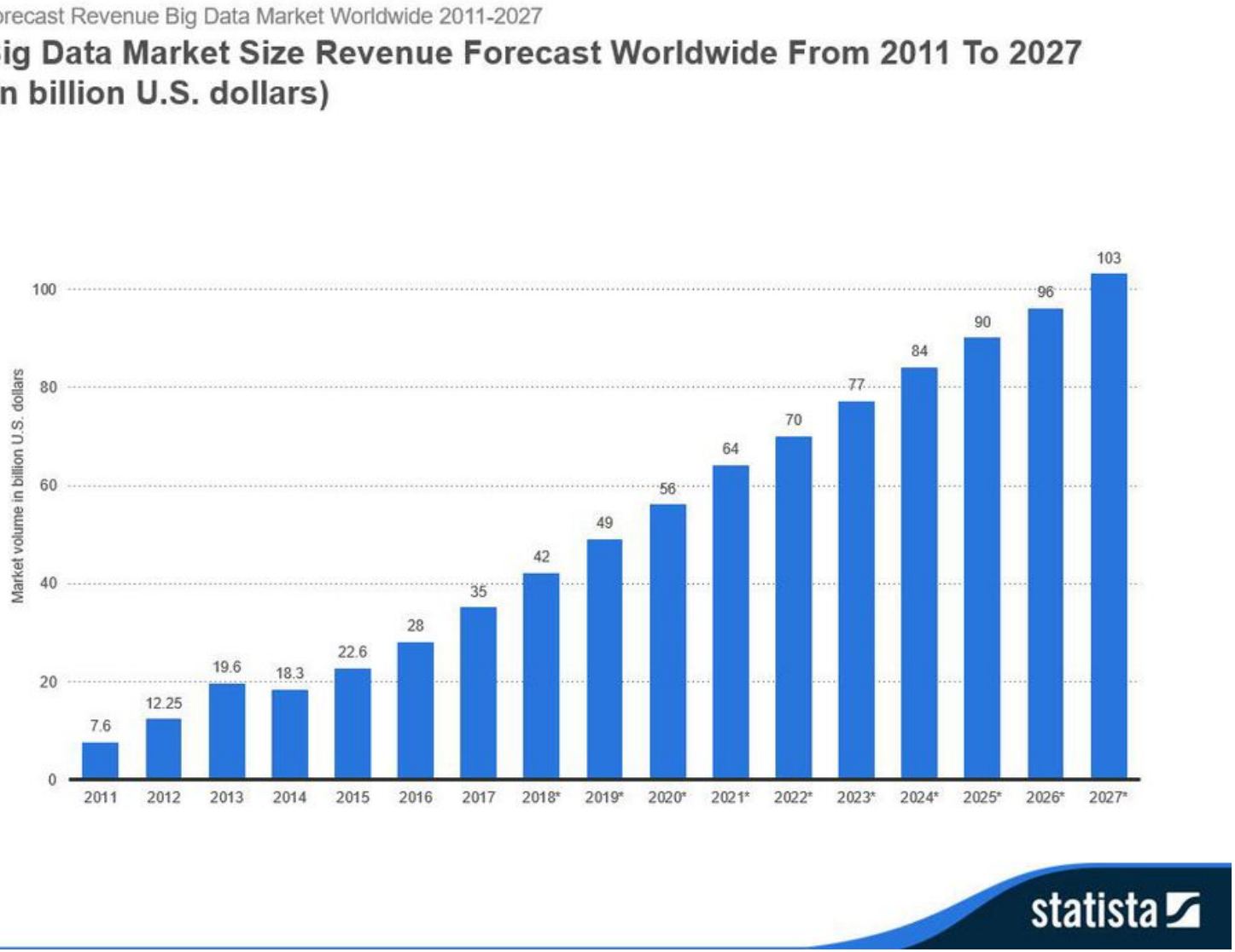




Big Data – future landscape

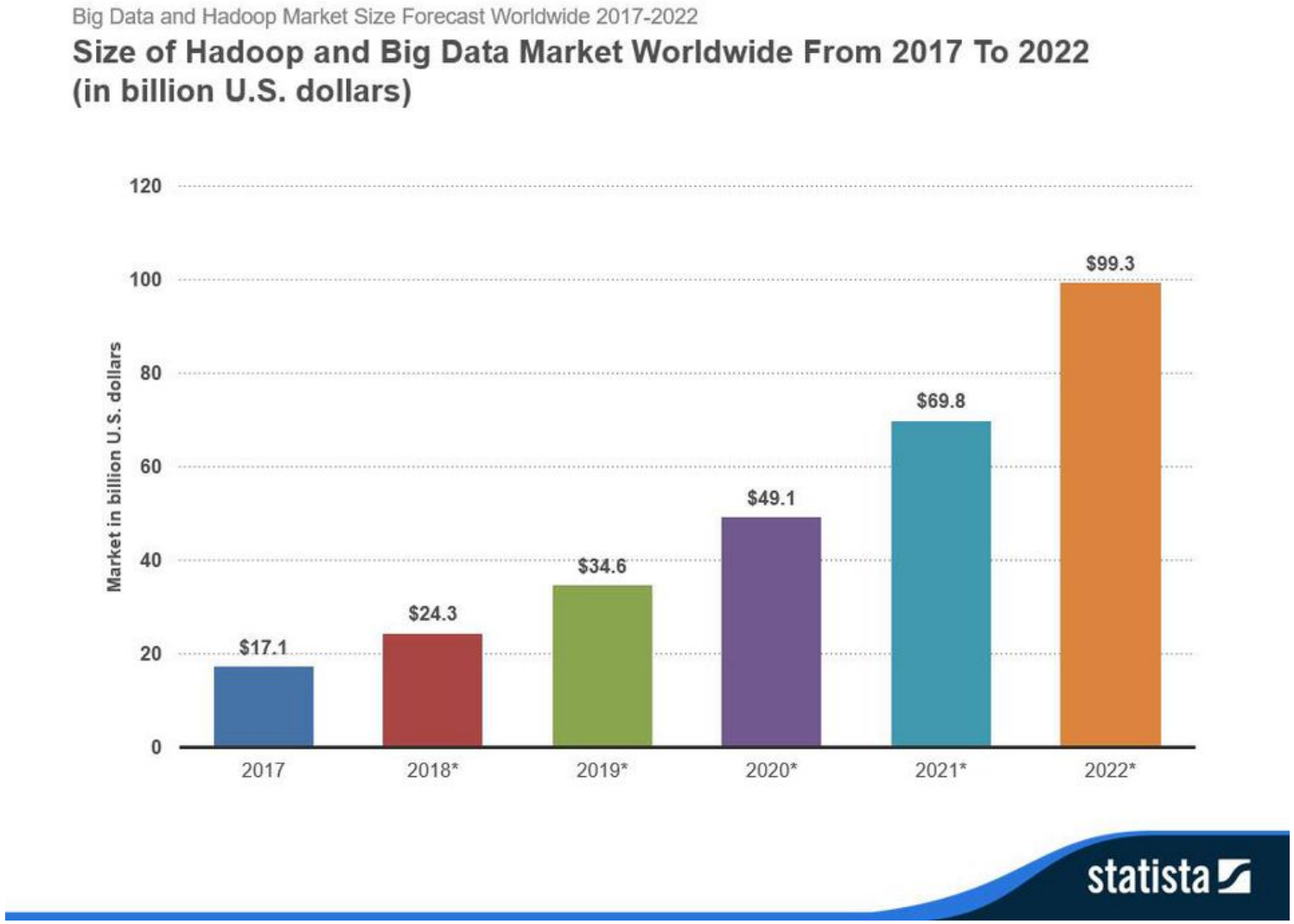
Ref: Analytics Comes of Age, published in January 2018 (PDF, 100 pp., no opt-in).

Big Data – future landscape



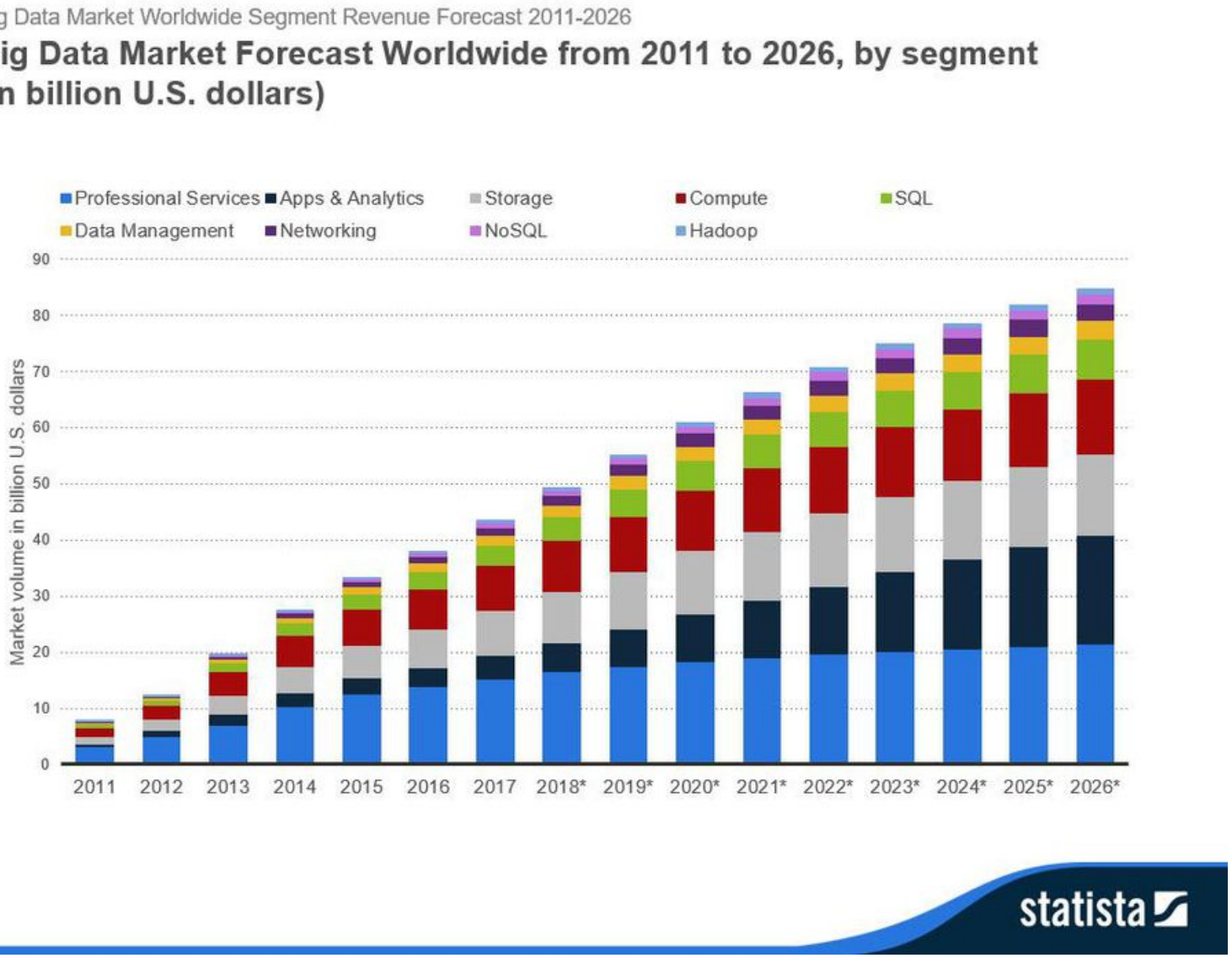
Worldwide Big Data market revenues for software and services are projected to increase from $56B in 2020 to $103B in 2027

Big Data – future landscape



The Hadoop and Big Data Market are projected to grow from $49.1B in 2017 to $99.31B in 2022

Big Data – future landscape



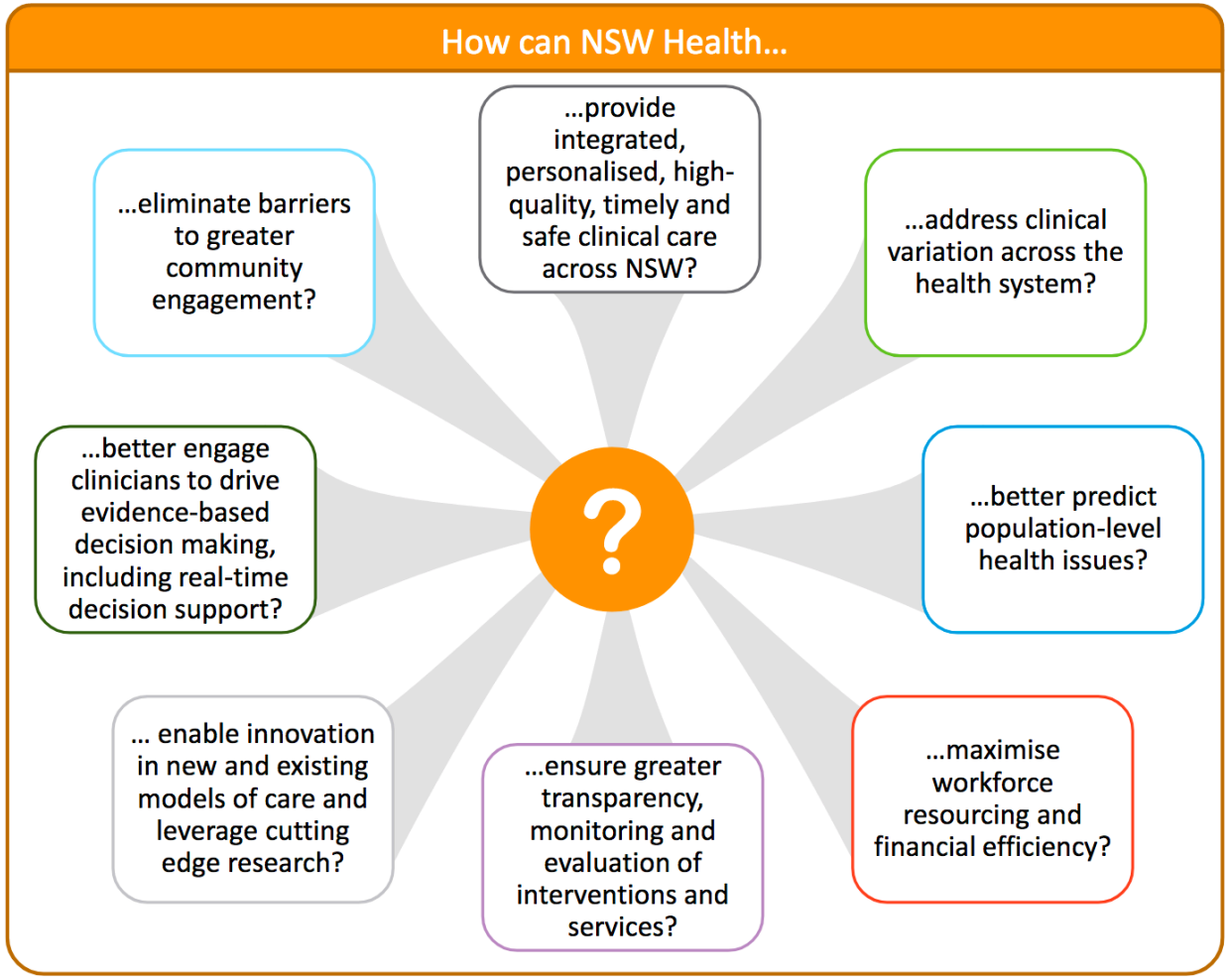
Big Data applications and analytics is projected to grow from $5.3B in 2018 to $19.4B in 2026

Data Governance Framework

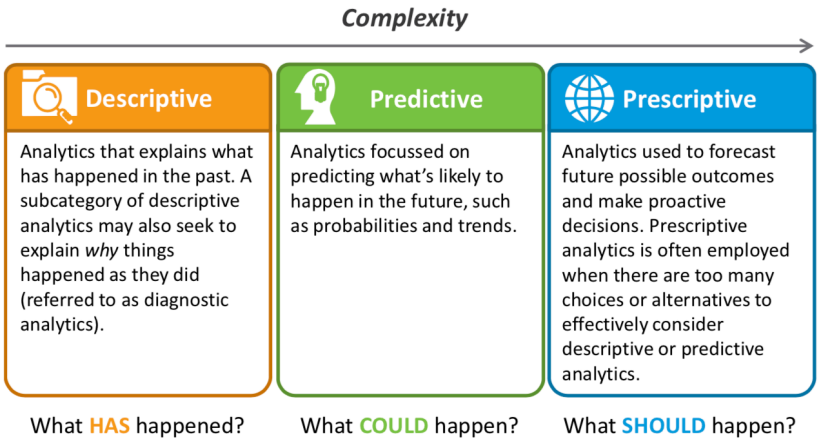
- outlines the roles and responsibilities involved in data governance and the structures
- ensure effective and consistent management of the data assets
- facilitates data quality and comprehensiveness, appropriate access to data, information security, and standardisation of concepts
- outlines the essential components of data governance, including description of the roles of
  - Data Sponsor
  - Data Custodian
  - Data Steward

https://www1.health.nsw.gov.au/pds/ActivePDSDocuments/GL2019_002.pdf

Data Analytics Framework

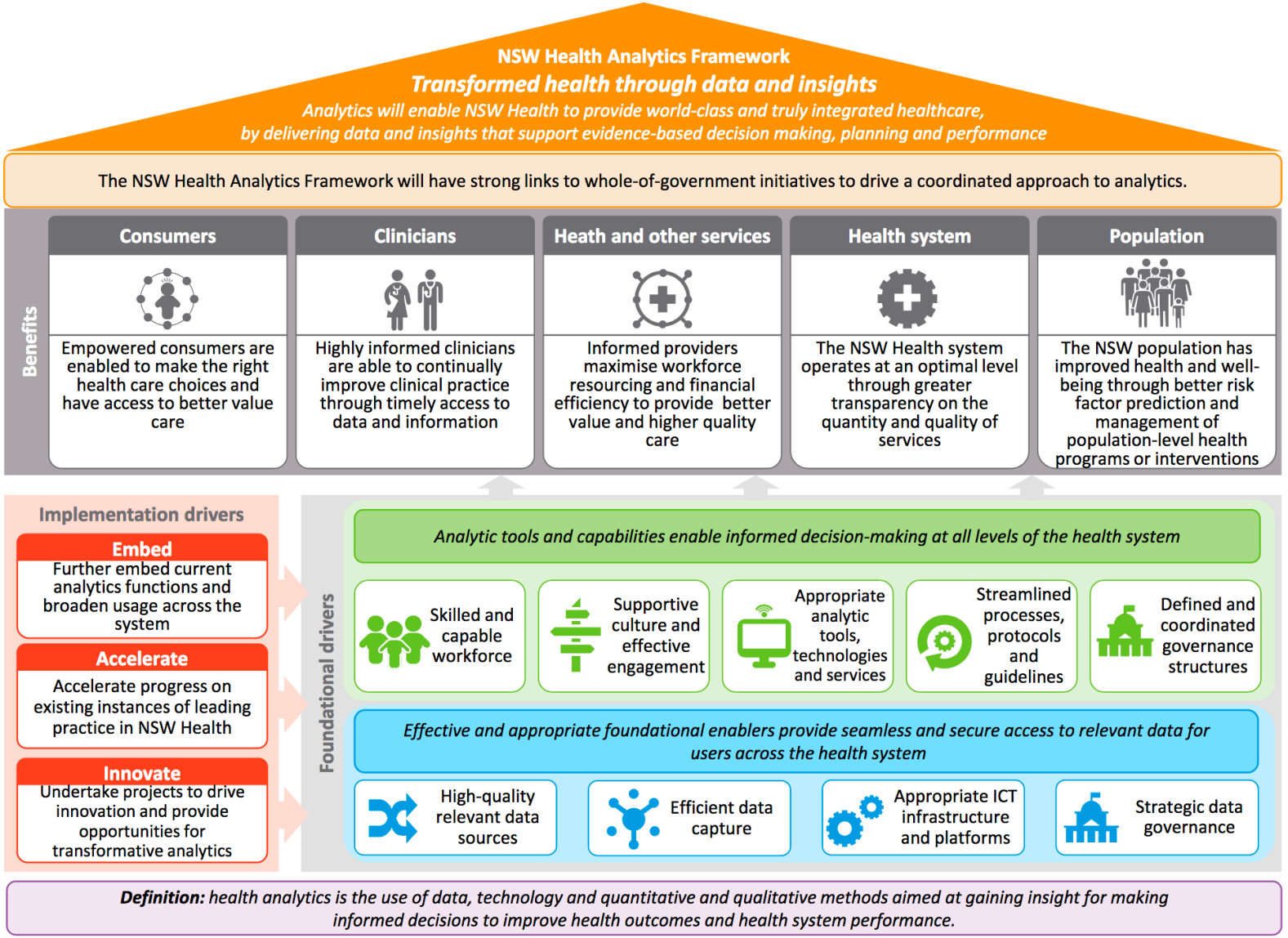


Types of analytics function



Ref: http://www.ehealth.nsw.gov.au/__data/assets/pdf_file/0020/303752/NSW_Health_Analytics_Framework.pdf

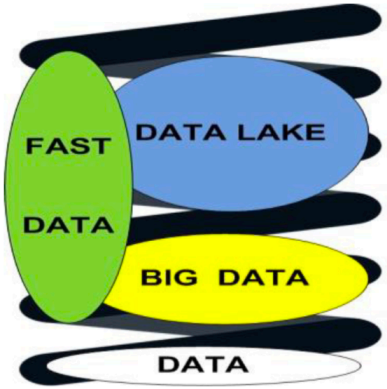
Data Analytics Framework



Ref. http://www.ehealth.nsw.gov.au/__data/assets/pdf_file/0020/303752/NSW_Health_Analytics_Framework.pdf

Data processing in enterprises

Purpose	Problem Space	Technology Characteristics
online transactional processing (OLTP)	Database transactions, Rollback, SOA, Web Services	RDBMS, Write and Lock intensive
online analytical processing (OLAP)	consolidation (roll-up), drill-down, and slicing and dicing; ETL (extract, transform, load) tool, data warehouse	In memory processing, data CUBE, ad hoc query, BI, Read intensive
Big Data	Four "V": volume, velocity, variety, veracity/variability Data Mining	Indexing, HDFS, Map Reduced, Batch Processing, Stream processing
Data Lake/Data Hub	Both OLAP and OLTP raw data, flat architecture, central data repository, multi-tier approach	Metadata tags, linked data, deep learning
Fast Data	Big Data analysis in real-time	Low latency, high input/output capability



Interrelation between big data, fast data and data lake concepts

Data processing techniques

JAMES COOK UNIVERSITY AUSTRALIA

Type	Problem Space	Example, Technology
Batch processing	sorted into groups to allow for efficient and sequential processing	ATMs are good examples
Real-time data processing	respond almost immediately to various signals to acquire and process information	Online trading/banking, many military applications, IoT
Multi processing	support more than one processor at the same time	multi-core, threading
Distributed processing	support more than one computer at the same time	Network, Hadoop
In memory processing	Effective use of memory and data representation	Compression, encoding, OLAP cube, apache spark
Parallel processing HW	Pipeline parallelism, Data parallelism	https://www.coursera.org/lecture/parallel-programming-in-java/4-4-pipeline-parallelism-9OMoh
Parallel processing SW	Shared memory, message passing	MAP Reduce, RabbitMQ

Dictionary encoding

JAMES COOK UNIVERSITY AUSTRALIA

Given a table (10 million rows) with the following columns:

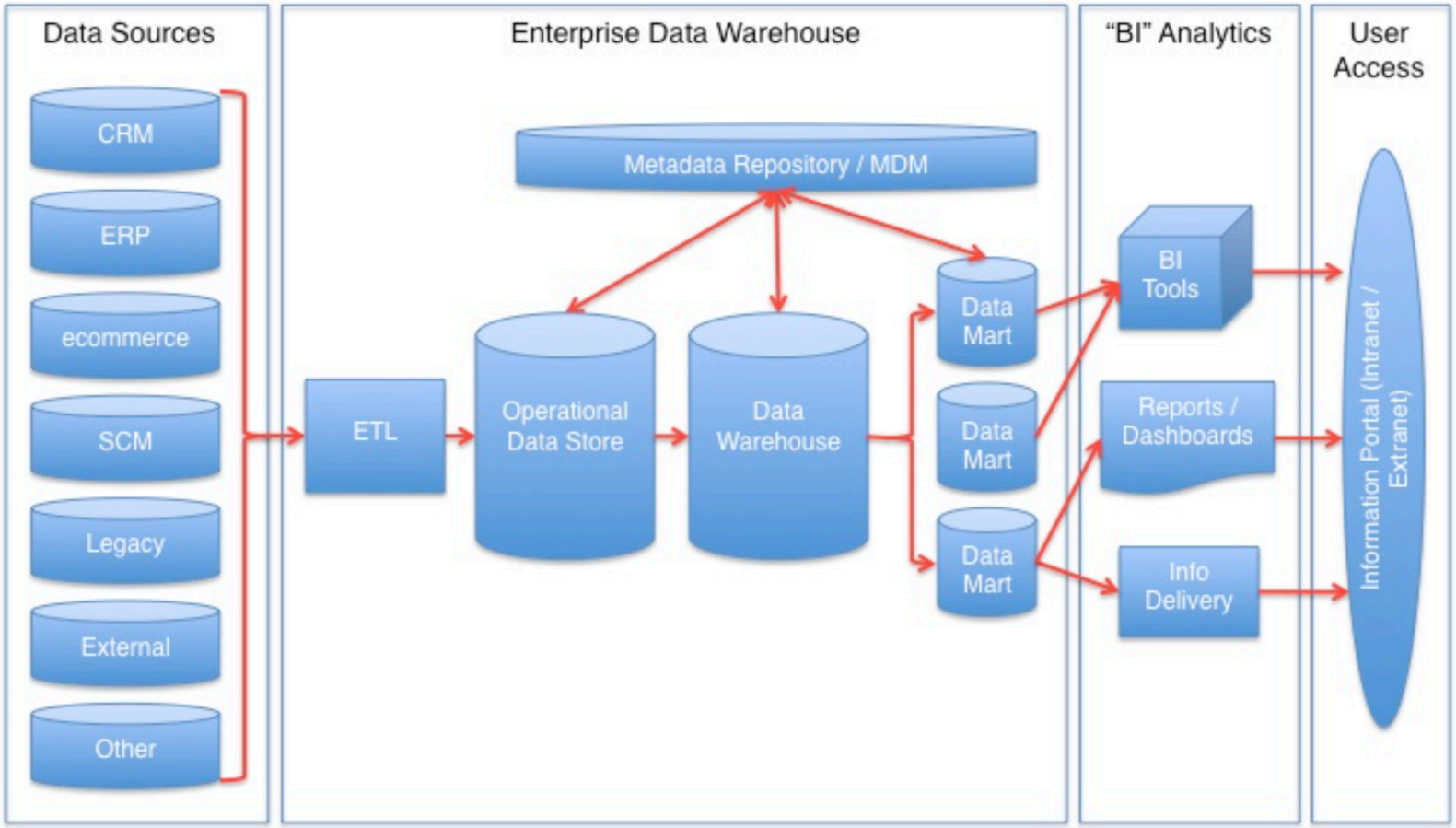
- EmployeeID (4 bytes, 10 million distinct values)
- Surname (49 bytes, 80000 distinct values)
- City (20 bytes, 40000 distinct values)
- Age (1 byte, 240 distinct values)
- Gender (1 byte, 3 distinct values)

Without encoding:
For a record it takes: 4 + 49 + 20 + 1 + 1 = 75 bytes
10 mil will take: 75bytes * 10 mil = 0.70 GB [1GB = 1024 MB]

Dictionary encoding applied:
Surname: log2(80000) = 17bits
City: log2(40000) = 16bits
Age: log2(240) = 8bits
Gender: log2(3) = 2bits
For one records it takes: 4*8+17+16+8+2 = 75bits
10 mil will take: 0.09GB

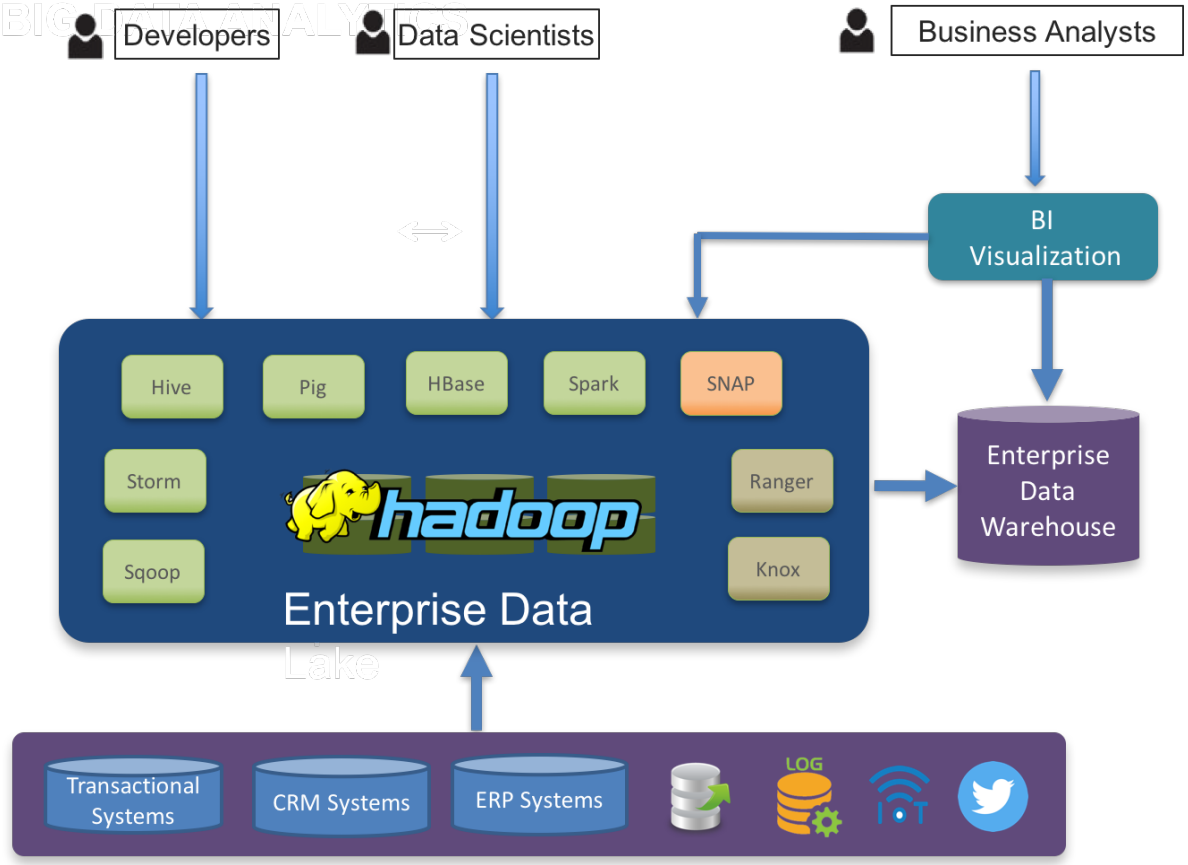
Compression ratio: 0.70/0.09 = 7.78

Data storage technologies



- Data warehouses are more structured in the delivery of information and subsequent data modelling, consequently they are used more frequently for decision support for past, present, and future
- Good for BI Reporting

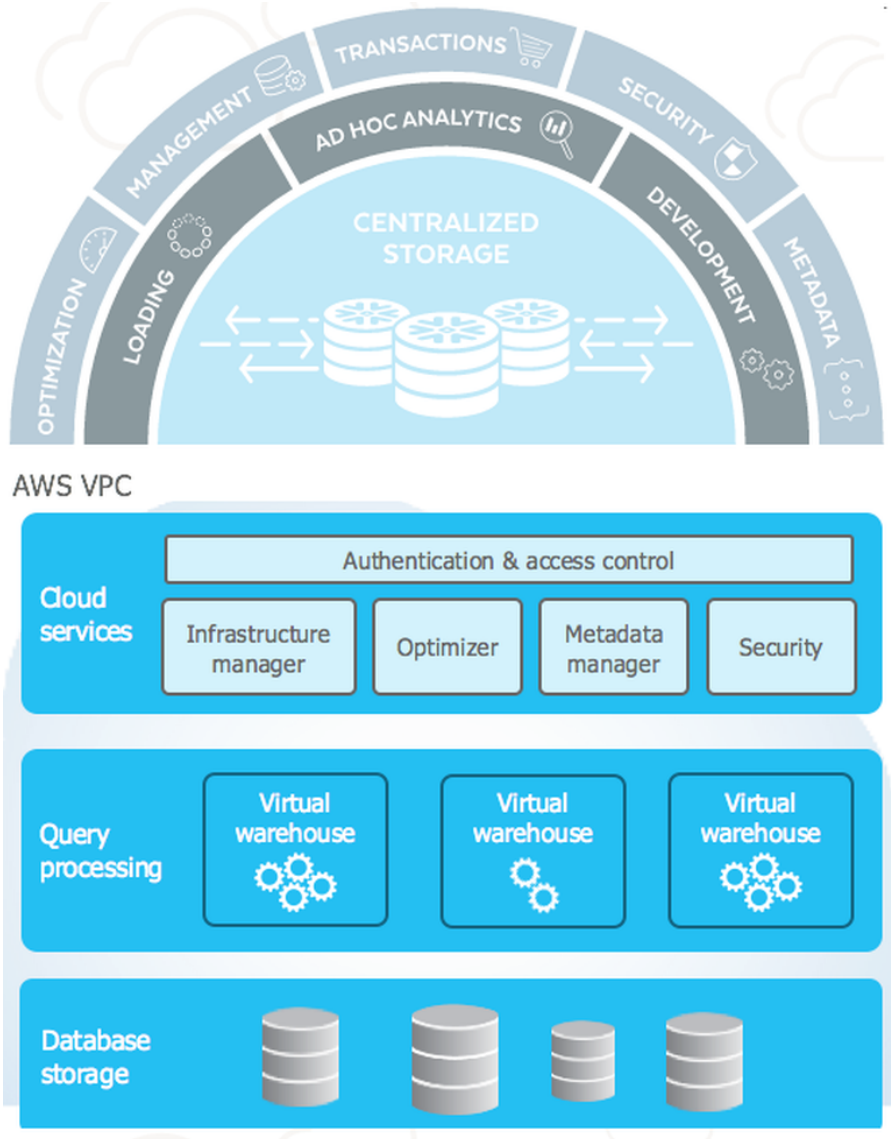
Data storage technologies

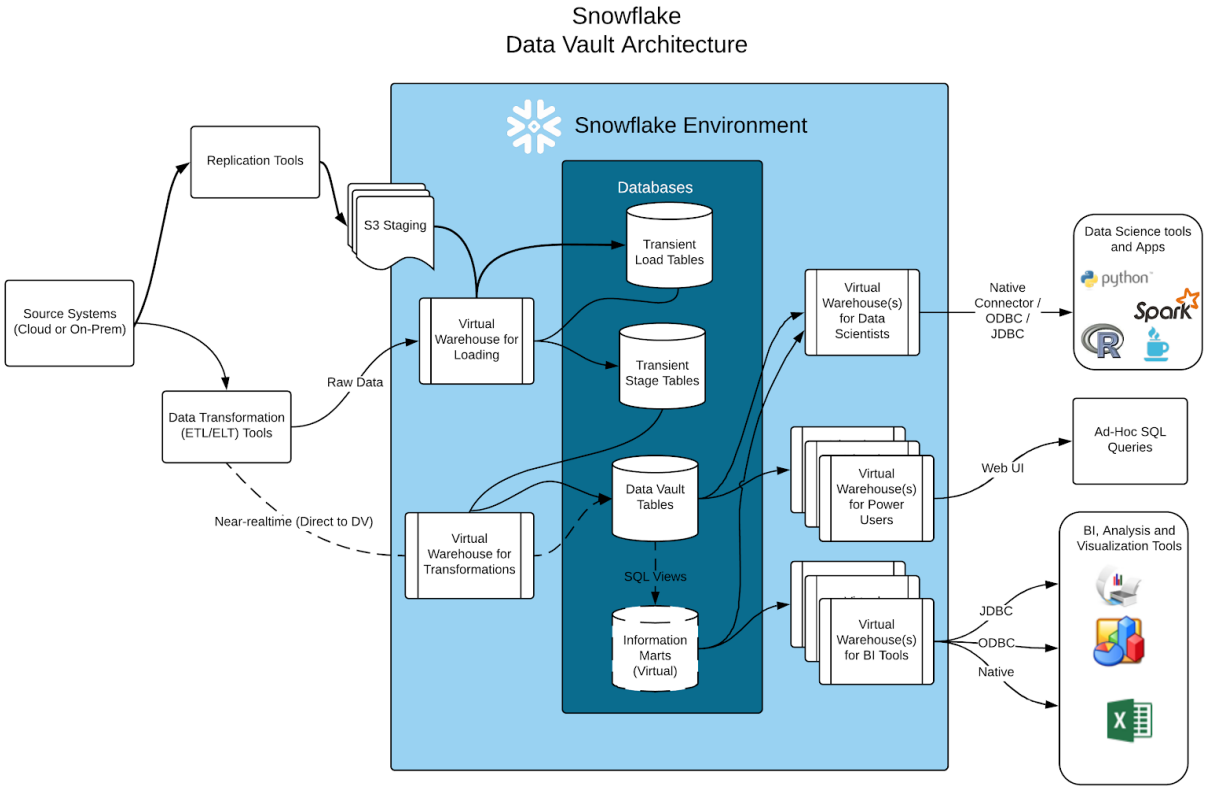


Data lakes are less well structured, ideal for big data, more flexible for use and more granular for querying

Not good for BI Reporting

Data warehouse on the cloud





- multi-cluster, shared data architecture, scale, elasticity, and concurrency → high performance

https://www.snowflake.com/product/architecture/