Introduction to Data Science, Data Analytic and Big Data

1


01 – Introduction to Data


02 – Data Science for Business


03 – Big Data Fundamentals


04 – Big Data Sources


05 – Introduction to Data Analytics


06 – Understanding Machine Learning

2

01 - Introduction to Data

3

Data is a collection of facts such as numbers, descriptions, and observations used in decision making.

What is data?

4

Data Categories

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

5

- A transactional system is often what most people consider the primary function of business computing.
- **A transactional system records transactions.**
- A transaction could be financial, such as the movement of money between accounts in a banking system, or it might be part of a retail system, tracking payments for goods and services from customers.

What is a Transactional System?

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

6

Customer		
CustomerID	CustomerName	CustomerPhone

Account	
CustomerID	Balance
5558	1000
6023	1500

Orders		
OrderID	CustomerID	OrderDate

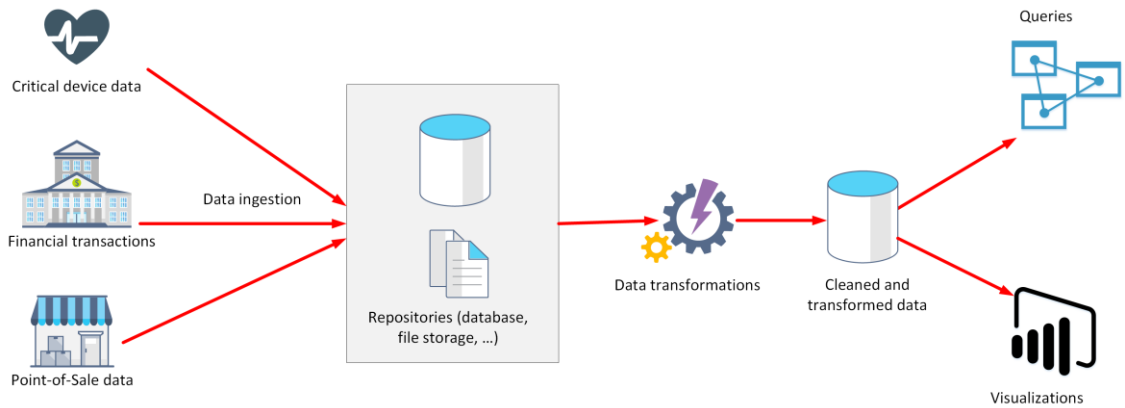
Transfers					
TransactionID	FromAccount	ToAccount	Transaction Amount	OrderDate	TransactionDescription
982801	6023	5558	500	DD/MM/YY	Transfer 500 from account 6023 to account

```
BEGIN TRANSACTION
UPDATE Account
SET Balance = Balance -500
WHERE CustomerID=6023;
UPDATE Account
SET Balance = Balance +500
WHERE CustomerID=5558;
INSERT INTO Transfers (Fromaccount, ToAccount, TransactionAmount, TransactionDescription)
VALUES (6023,5558,500, 'Transfer 500 from account 6023 to account 5558)
COMMIT TRANSACTIONTransaction
```

Transactional workloads

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

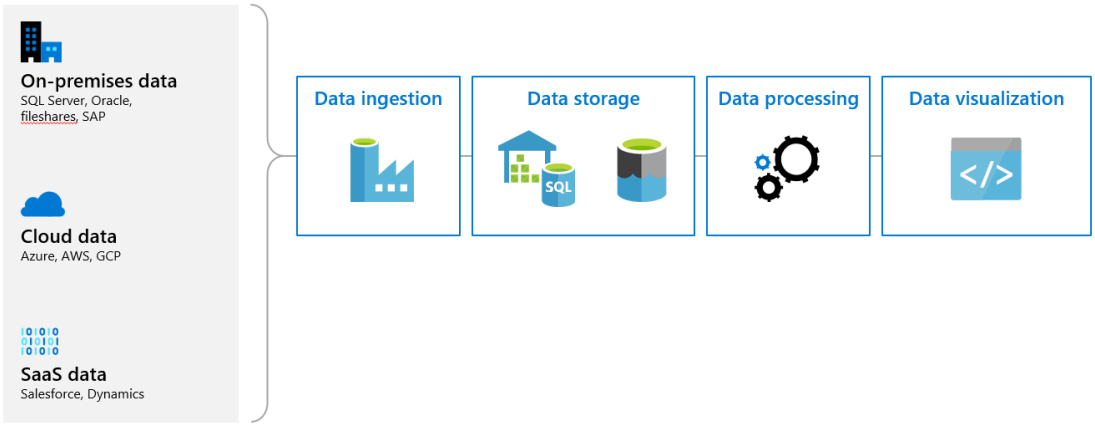
7



What is an Analytical System?

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

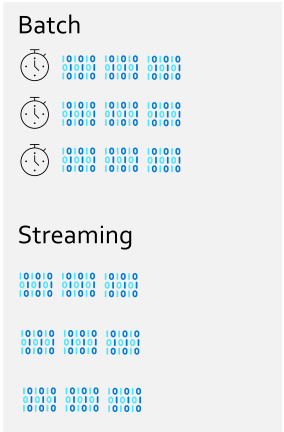
8

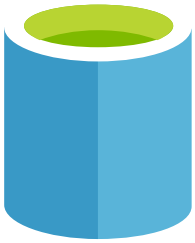


What is an Analytical System?

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

9





Batch Data / Streaming Data

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

10

CustomerID	FirstName	LastName
1	Jay	Adams
2	Donna	Carreras
3	Linda	Burnett
4	Frances	Adams

CustomerID	AddressID
1	A
2	B
3	C
4	A

AddressID	LineNumber	Text
A	1	12
A	2	Park Street
A	3	Some City
B	1	The Big House
B	2	High Road
B	3	Another City
B	4	90210
C	1	Freepost
C	2	AAA 123

Relational Data

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

11

- All data is tabular. Entities are modeled as tables, each instance of an entity is a row in the table, and each property is defined as a column.
- All rows in the same table have the same set of columns.
- A table can contain any number of rows.

Characteristics of a Relational Database

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

12

- A primary key uniquely identifies each row in a table. No two rows can share the same primary key.
- A foreign key references rows in another, related table. For each value in the foreign key column, there should be a row with the same value in the corresponding primary key column in the other table.

Characteristics of a Relational Database

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

13

- SQL is a standard language for use with relational databases
- SQL standards are maintained by ANSI and ISO
- Proprietary RDBMS systems have their own extensions of SQL such as T-SQL, PL/SQL, pgSQL

Query - SQL

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

14

DML	DDL	DCL
Data Manipulation Language	Data Definition Language	Data Control Language
Used to query and manipulate data	Used to define database objects	Used to manage security permissions
SELECT, INSERT, UPDATE, DELETE	CREATE, ALTER, DROP	GRANT, REVOKE, DENY

Query - SQL

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

15

Statement	Description
SELECT	Select/read from a table
INSERT	Insert new rows in a table
UPDATE	Edit/Update existing rows in a table
DELETE	Delete existing rows in a table

Query - SQL

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

16

```sql
SELECT EmployeeId, YEAR(OrderDate) AS OrderYear
FROM Sales.Orders
WHERE CustomerId = 71
GROUP BY EmployeeId, YEAR(OrderDate)
HAVING COUNT(*) > 1
ORDER BY EmployeeId, OrderYear;
```

Query - SQL

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

17

Statement	Description
CREATE	Create a new object in the database, such as a table or a view
INSERT	Modify the structure of an object. For instance, altering a table to add a new column
UPDATE	Remove an object from the database
DELETE	Rename an existing object

Query - SQL

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

18

```
CREATE TABLE Mytable
(Mycolumn1 int NOT NULL PRIMARY KEY, Mycolumn2 VARCHAR(50) NOT
NULL , Mycolumn2 VARCHAR(10) NOT NULL)
```

Query - SQL

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

19

Customers		
CustomerID	CustomerName	CustomerPhone
100	Muisto Linna	*XXX-XXX-XXXX*
101	Noam Maoz	*XXX-XXX-XXXX*
102	Vanja Matkovic	*XXX-XXX-XXXX*
103	Qamar Mounir	*XXX-XXX-XXXX*
104	Zhenis Omar	*XXX-XXX-XXXX*
105	Claude Paulet	*XXX-XXX-XXXX*
106	Alex Pettersen	*XXX-XXX-XXXX*
107	Francis Ribeiro	*XXX-XXX-XXXX*

Tables

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

20

Customers		
CustomerID	CustomerName	CustomerPhone
100	Muisto Linna	*XXX-XXX-XXXX*
101	Noam Maoz	*XXX-XXX-XXXX*
102	Vanja Matkovic	*XXX-XXX-XXXX*
103	Qamar Mounir	*XXX-XXX-XXXX*
104	Zhenis Omar	*XXX-XXX-XXXX*
105	Claude Paulet	*XXX-XXX-XXXX*
106	Alex Pettersen	*XXX-XXX-XXXX*

Orders		
OrderID	CustomerName	CustomerPhone
AD100	Noa oz	*XXX-XXX-XXXX*
AD101	Noam	*XXX-XXX-XXXX*
AD102	Noam Ma	*XXX-XXX-XXXX*
AX103	Qamar M	*XXX-XXX-XXXX*
AS104	Qam nir	*XXX-XXX-XXXX*
AR105	Claude Paulet	*XXX-XXX-XXXX*
MK106	Muisto Linna	*XXX-XXX-XXXX*

Data is normalized to:

Reduce storage	Avoid data duplication	Improve data quality

Normalization

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

21

Customers		
CustomerID	CustomerName	CustomerPhone
100	Muisto Linna	*XXX-XXX-XXXX*
101	Noam Maoz	*XXX-XXX-XXXX*
102	Vanja Matkovic	*XXX-XXX-XXXX*
103	Qamar Mounir	*XXX-XXX-XXXX*
104	Zhenis Omar	*XXX-XXX-XXXX*
105	Claude Paulet	*XXX-XXX-XXXX*
106	Alex Pettersen	*XXX-XXX-XXXX*

Orders		
OrderID	CustomerID	SalesPersonID
AD100	101	*200*
AD101	101	200
AD102	101	200
AX103	10	201
AS104	103	201
AR105	105	200
MK106	105	201

In a normalized database schema:

Primary Keys and Foreign keys are used to define relationships	No data duplication exists (other than key values in 3rd Normal Form (3NF)	Data is retrieved by joining tables together in a query

Relations

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

22

Customers		
CustomerID	CustomerName	CustomerPhone
100	Muisto Linna	*XXX-XXX-XXXX*
101	Noam Maoz	*XXX-XXX-XXXX*
102	Vanja Matkovic	*XXX-XXX-XXXX*
103	Qamar Mounir	*XXX-XXX-XXXX*
104	Zhenis Omar	*XXX-XXX-XXXX*
105	Claude Paulet	*XXX-XXX-XXXX*
106	Alex Pettersen	*XXX-XXX-XXXX*

IDX-CustomerRegion	
CustomerID	Region
100	France
101	Brazil
102	Croatia
103	Jordan
104	Spain
105	France
106	USA

An index:

- Optimizes search queries for faster data retrieval
- Reduces the amount of data pages that need to be read to retrieve the data in a SQL Statement
- Data is retrieved by joining tables together in a query

Indexes

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

23

Customers		
CustomerID	CustomerName	CustomerPhone
100	Muisto Linna	*XXX-XXX-XXXX*
101	Noam Maoz	*XXX-XXX-XXXX*
102	Vanja Matkovic	*XXX-XXX-XXXX*
103	Qamar Mounir	*XXX-XXX-XXXX*
104	Zhenis Omar	*XXX-XXX-XXXX*
105	Claude Paulet	*XXX-XXX-XXXX*
106	Alex Pettersen	*XXX-XXX-XXXX*

Orders		
OrderID	CustomerID	SalesPersonID
AD100	101	*200*
AD101	101	200
AD102	101	200
AX103	103	201
AS104	103	201
AR105	105	200
MK106	105	201
DB205	100	205

```
Create the definition of a view:

CREATE VIEW vw_customerorders AS

SELECT Customers.CustomerID, Customers.CustomerName, Orders.OrderID FROM Customers JOIN Orders on Customers.CustomerID = Orders.CustomerID

Retrieve the orders placed by customer 102 using the view:

SELECT CustomerName, OrderID from vw_customerorders WHERE CustomerID=102
```

A view is a virtual table based on the result set of query:

- Views are created to simplify the query
- Combine relational data into a single pane view

View

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

24

Relational Database Management System - RDBMS

25

```
## Document for Jay Adams ##
{
  "customerID": "1",
  "name":
  {
    "firstname": "Jay",
    "lastname": "Adams"
  },
  "address":
  {
    "number": "12",
    "street": "Park Street",
    "city": "Some City",
  }
}
```

```
## Document for Frances Adams ##
{
  "customerID": "4",
  "name":
  {
    "firstname": "Francis",
    "lastname": "Adams"
  },
  "address":
  {
    "number": "12",
    "street": "Park Street",
    "city": "Some City",
  }
}
```

Non-Relational Data

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

26

```
## Customer 1 ID: 1
Name: Mark Hanson
Telephone: [ Home: 1-999-9999999, Business: 1-888-8888888, Cell: 1-777- 7777777 ]
Address: [ Home: 121 Main Street, Some City, NY, 10110,
           Business: 87 Big Building, Some City, NY, 10111 ]
## Customer 2 ID: 2
Title: Mr
Name: Jeff Hay
Telephone: [ Home: 0044-1999-333333, Mobile: 0044-17545-444444 ]
Address: [ UK: 86 High Street, Some Town, A County, GL8888, UK,
           US: 777 7th Street, Another City, CA, 90111 ]
```

Non-relational collections can have:

Multiple entities in the same collection or container with different fields	Have a different, non-tabular schema	Are often defined by labeling each field with the name it represents

Non-Relational Data

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

27

- IoT and Telematics
- Retail and Marketing
- Gaming
- Web and Mobile

Identify non-relational database use cases

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

28

```
{"latitude":37.8267,"longitude":-122.4233,"timezone":"America/Los_Angeles","currently":{"time":1598191217,"summary":"Partly Cloudy","icon":"partly-
cloudy-
day","nearestStormDistance":5,"nearestStormBearing":58,"precipIntensity":0,"precipProbability":0,"temperature":58.63,"apparentTemperature":58.63,"dew
Point":52.42,"humidity":0.8,"pressure":1011.8,"windSpeed":5.08,"windGust":7.73,"windBearing":210,"cloudCover":0.54,"uvIndex":0,"visibility":9.933,"oz
one":291.2},"minutely":{"summary":"Partly cloudy for the hour.","icon":"partly-cloudy-day","data":
[{"time":1598191200,"precipIntensity":0,"precipProbability":0},{"time":1598191260,"precipIntensity":0,"precipProbability":0},
{"time":1598191320,"precipIntensity":0,"precipProbability":0},{"time":1598191380,"precipIntensity":0,"precipProbability":0},
{"time":1598191440,"precipIntensity":0,"precipProbability":0},{"time":1598191500,"precipIntensity":0,"precipProbability":0},
{"time":1598191560,"precipIntensity":0,"precipProbability":0},{"time":1598191620,"precipIntensity":0,"precipProbability":0},
{"time":1598191680,"precipIntensity":0,"precipProbability":0},{"time":1598191740,"precipIntensity":0,"precipProbability":0},
{"time":1598191800,"precipIntensity":0,"precipProbability":0},{"time":1598191860,"precipIntensity":0,"precipProbability":0},
{"time":1598191920,"precipIntensity":0,"precipProbability":0},{"time":1598191980,"precipIntensity":0,"precipProbability":0},
{"time":1598192040,"precipIntensity":0,"precipProbability":0},{"time":1598192100,"precipIntensity":0,"precipProbability":0},
{"time":1598192160,"precipIntensity":0,"precipProbability":0},{"time":1598192220,"precipIntensity":0,"precipProbability":0},
{"time":1598192280,"precipIntensity":0,"precipProbability":0},{"time":1598192340,"precipIntensity":0,"precipProbability":0},
{"time":1598192400,"precipIntensity":0.0026,"precipIntensityError":0.0004,"precipProbability":0.01,"precipType":"rain"},
{"time":1598192460,"precipIntensity":0.0026,"precipIntensityError":0.0005,"precipProbability":0.01,"precipType":"rain"},
{"time":1598192520,"precipIntensity":0,"precipProbability":0},
{"time":1598192580,"precipIntensity":0.0026,"precipIntensityError":0.0005,"precipProbability":0.01,"precipType":"rain"},
{"time":1598192640,"precipIntensity":0.0026,"precipIntensityError":0.0005,"precipProbability":0.01,"precipType":"rain"},
{"time":1598192700,"precipIntensity":0.0027,"precipIntensityError":0.0005,"precipProbability":0.01,"precipType":"rain"},
{"time":1598192760,"precipIntensity":0.0027,"precipIntensityError":0.0005,"precipProbability":0.02,"precipType":"rain"},
{"time":1598192820,"precipIntensity":0.0026,"precipIntensityError":0.0005,"precipProbability":0.02,"precipType":"rain"},
{"time":1598192880,"precipIntensity":0,"precipProbability":0},
```


Non-Relational Database use case

https://api.darksky.net/

29

Open API :] สำหรับนักพัฒนา

แสดงค่าประจำวัน :
//covid19.th-stat.com/api/open/today

ข้อมูลสรุปตามช่วงเวลา [เริ่มตั้งแต่วันที่ 01/01/20] :
//covid19.th-stat.com/api/open/timeline

ข้อมูลแต่ละเคส :
//covid19.th-stat.com/api/open/cases

ข้อมูลสรุปจากเคส :
//covid19.th-stat.com/api/open/cases/sum

แจ้งเตือนพื้นที่ตามคำประกาศ :
//covid19.th-stat.com/api/open/area

กรมควบคุมโรค
DEPARTMENT OF DISEASE CONTROL

Non-Relational Database use case

30

	Schema	Data relationships	Examples
Structured data	Adheres to a schema, with the same data fields or properties.	Storable in relational database tables, with rows and columns.	Sensor data and financial data.
Semi-structured data	Has an ad hoc schema with less organized fields and properties.	Non-relational or NoSQL data, not storable in tables, rows and column.	Books, blogs, JSON, HTML documents.
Unstructured data	Has no designated schema or data structure.	Non-relational or blob data, with no restrictions on the kinds of data blobs contain.	PDFs, JPGs, videos.

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

31

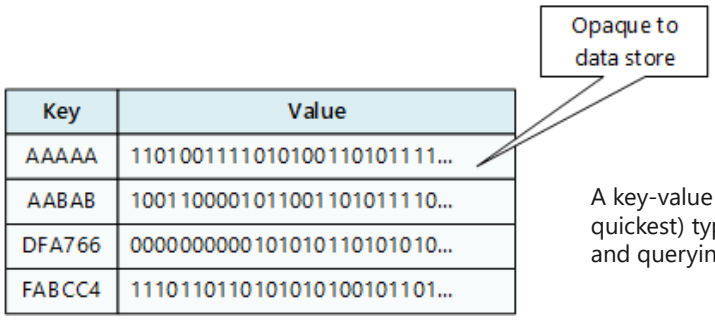
- You might see the term *NoSQL* when reading about non-relational databases.
- NoSQL is a rather loose term that simply means non-relational.
- NoSQL (non-relational) databases generally fall into four categories:

  - key-value stores
  - document databases
  - column family databases
  - graph databases.

What is NoSQL?

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

32

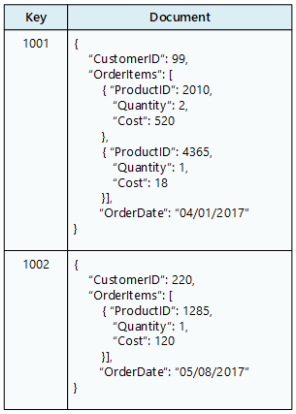


A key-value store is the simplest (and often quickest) type of NoSQL database for inserting and querying data.

Key-Value Stores

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

33



A document database represents the opposite end of the NoSQL spectrum from a key-value store. In a document database, each document has a unique ID, but the fields in the documents are transparent to the database management system. Document databases typically store data in JSON format,

Document Databases

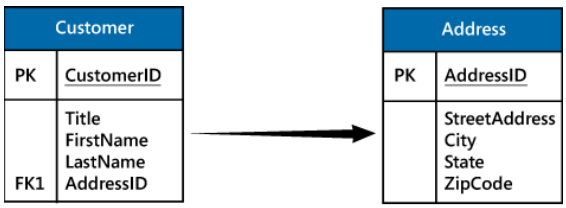
Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

34

RDBMS	MongoDB
Database	Database
Table	Collection
Tuple/Row	Document
column	Field
Table Join	Embedded Documents
Primary Key	Primary Key (Default key _id provided by mongodb itself)

Document Database

35



RDBMS is Row-based oriented

Customer Table

CustomerID	Title	FirstName	LastName	AddressID
1	Mr	Mark	Hanson	500
2	Ms	Lisa	Andrews	501
3	Mr	Walter	Harp	500

Address Table

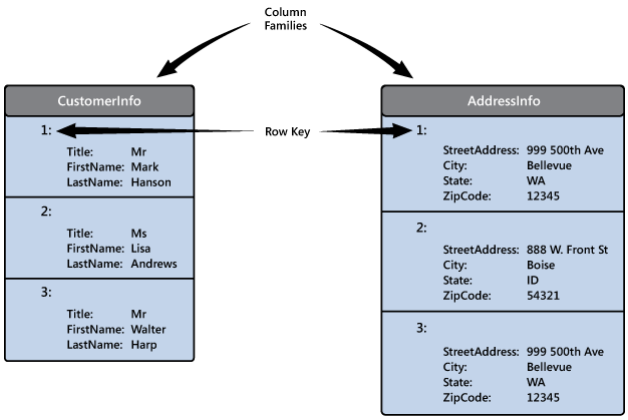
AddressID	StreetAddress	City	State	ZipCode
500	999 500th Ave	Bellevue	WA	12345
501	888 W. Front St	Boise	ID	54321

Column Family Databases

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

36

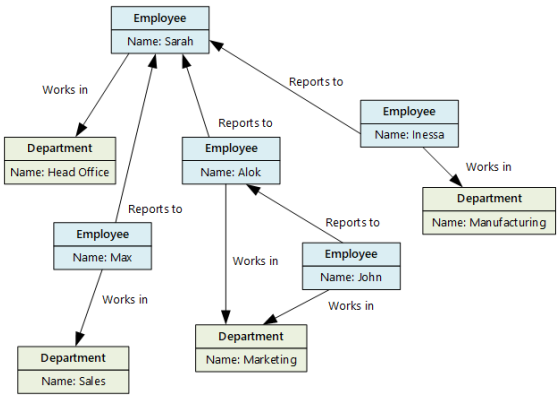
Row Key	Column Families	
CustomerID	CustomerInfo	AddressInfo
1	CustomerInfo:Title Mr CustomerInfo:FirstName Mark CustomerInfo:LastName Hanson	AddressInfo:StreetAddress 999 500th Ave AddressInfo:City Bellevue AddressInfo:State WA AddressInfo:ZipCode 12345
2	CustomerInfo:Title Ms CustomerInfo:FirstName Lisa CustomerInfo:LastName Andrews	AddressInfo:StreetAddress 888 W. Front St AddressInfo:City Boise AddressInfo:State ID AddressInfo:ZipCode 54321
3	CustomerInfo:Title Mr CustomerInfo:FirstName Walter CustomerInfo:LastName Harp	AddressInfo:StreetAddress 999 500th Ave AddressInfo:City Bellevue AddressInfo:State WA AddressInfo:ZipCode 12345



Column Family Databases

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

37



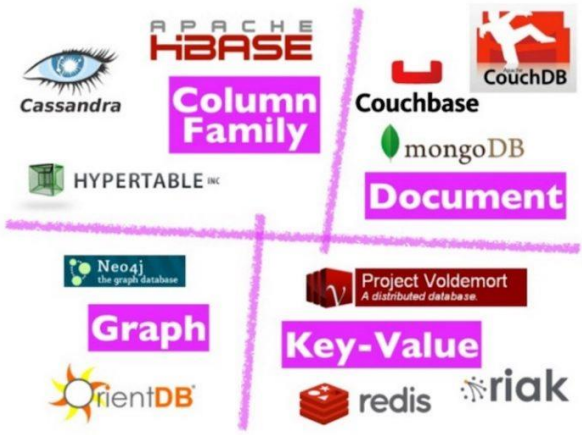
Graph databases enable you to store entities, but the main focus is on the relationships that these entities have with each other.

A graph database stores two types of information: nodes that you can think of as instances of entities, and edges, which specify the relationships between nodes.

Graph Databases

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

38



No-SQL Tools

Content Reference : www.somkiat.cc

39

Database Administrator	Data Engineer	Data Analyst
Database Management	Data Pipelines and processes	Provides insights into the data
Implements Data Security	Data Ingestion storage	Visual Reporting
Backups	Prepare data for Analytics	Modeling Data for Analysis
User Access	Prepare data for analytical processing	Combines data for visualization and analysis
Monitors performance		

Roles in Data

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

40

- Installing and upgrading the database server and application tools.
- Allocating system storage and planning storage requirements for the database system.
- Modifying the database structure, as necessary, from information given by application developers.
- Enrolling users and maintaining system security.

Database Administrator tasks and responsibilities

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

41

- Ensuring compliance with database vendor license agreement.
- Controlling and monitoring user access to the database.
- Monitoring and optimizing the performance of the database.
- Planning for backup and recovery of database information.
- Maintaining archived data.

Database Administrator tasks and responsibilities

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

42

- Backing up and restoring databases.
- Contacting database vendor for technical support.
- Generating various reports by querying from database as per need.
- Managing and monitoring data replication.
- Acting as liaison with users.

Database Administrator tasks and responsibilities

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

43

- Developing, constructing, testing, and maintaining databases and data structures.
- Aligning the data architecture with business requirements.
- Data acquisition.
- Developing processes for creating and retrieving information from data sets.
- Using programming languages and tools to examine the data.

Data Engineer tasks and responsibilities

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

44

- Identifying ways to improve data reliability, efficiency, and quality.
- Conducting research for industry and business questions.
- Deploying sophisticated analytics programs, machine learning, and statistical methods.
- Preparing data for predictive and prescriptive modeling.
- Using data to discover tasks that can be automated.

Data Engineer tasks and responsibilities

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

45

- Making large or complex data more accessible, understandable, and usable.
- Creating charts and graphs, histograms, geographical maps, and other visual models that help to explain the meaning of large volumes of data, and isolate areas of interest.
- Transforming, improving, and integrating data from many sources, depending on the business requirements.

Data Analyst tasks and responsibilities

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

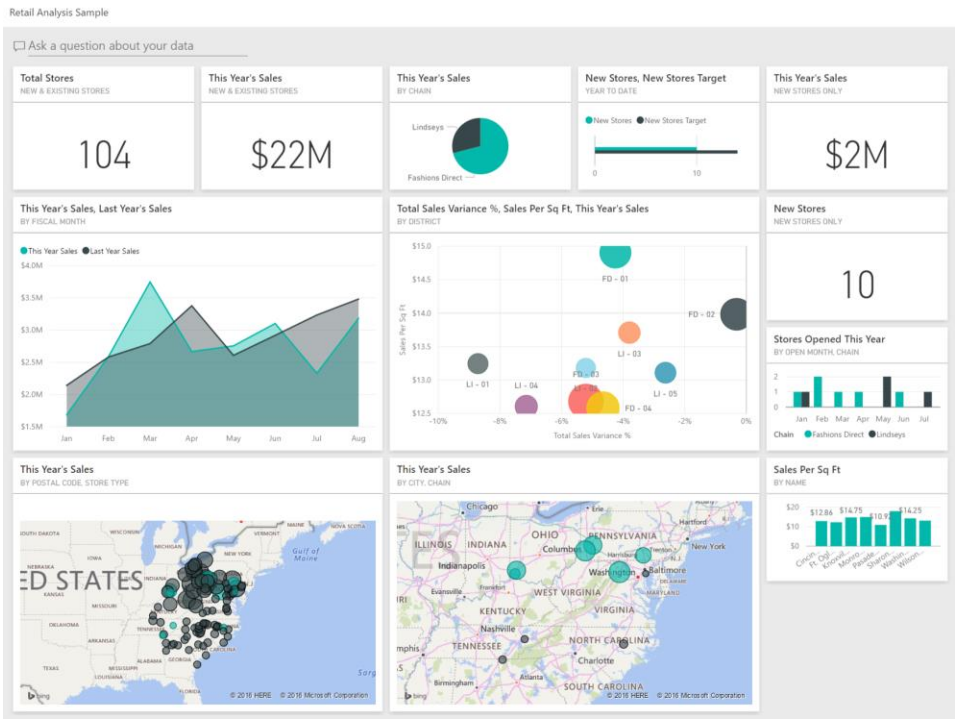
46

- Combining the data result sets across multiple sources. For example, combining sales data and weather data provides a useful insight into how weather influenced sales of certain products such as ice creams.
- Finding hidden patterns using data.
- Delivering information in a useful and appealing way to users by creating rich graphical dashboards and reports.

Data Analyst tasks and responsibilities

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

47



48

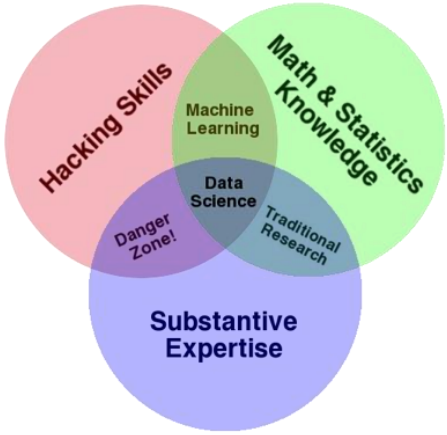
02 - Data Science for Business

49

- Data Science aims to derive **knowledge** from **big data, efficiently** and **intelligently.**
- Data Science encompasses the **set of activity, tools, and methods** that enable **data-driven activities** in science, business, medicine, and government.
- Data science is a multidisciplinary blend of **data inference, algorithm development, and technology** in order to solve analytically complex problems.

What is Data Science?

50



Data Science Venn Diagram (Drew Conway)

Content Reference : http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

51

Element	Databases	Data Science
Data Value	Precious	Cheap
Data Volume	Modest	Massive
Example	Bank records, Census Information, Medical records, ...	Online clicks, GPS logs, Social Media (Facebook, Tweets), ...
Priorities	Consistency, Error Recovery, Auditability	Speed, Availability, Query richness
Structured	Strongly (Schema)	Weakly or None (Text)

Data Science vs. Database

Content Reference : Veerasak Kritsanapraphan [Software Park Thailand]

52

Element	Databases	Data Science
Properties	Transactions, Atomicity, Consistency, Isolation, Durability	Consistency, Availability, Partition Tolerance, theorem (2/3), eventual consistency
Realizations	SQL (Oracle, SQL Server, MariaDB, …)	NoSQL (MongoDB, Apache Hbase, …)
Query	Querying the past	Querying the future

Data Science vs. Database

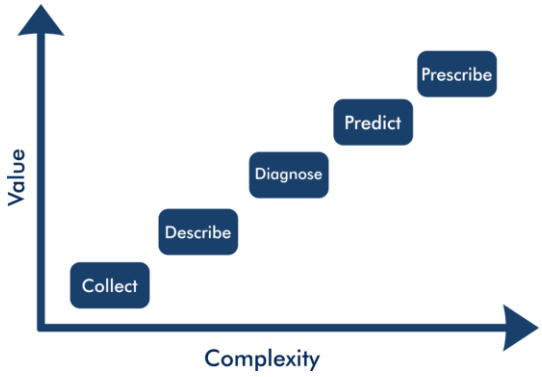
Content Reference : Veerasak Kritsanapraphan [Software Park Thailand]

53

- Cloud Computing
- Big Data
- Machine Learning
- Statistics and Probability
- Programming Language (R, Python)

Basic Components of Data Science

54

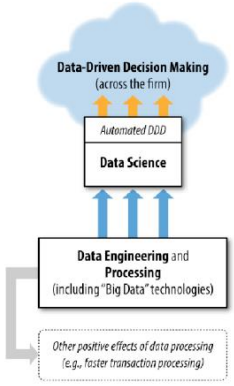


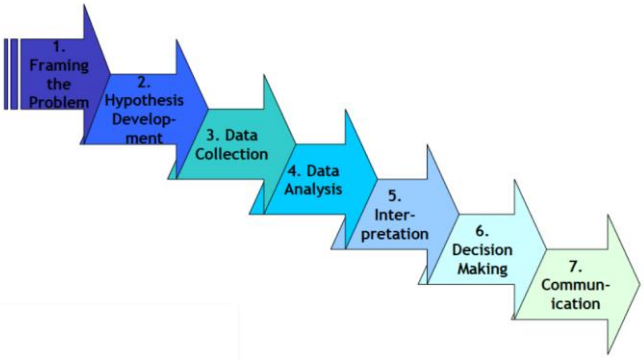
- Describing the data is the first step in extracting value.
- Descriptive statistics are the core of most business reporting and are an essential first step in analysing the data.
- Diagnostics or analysis is the core activity of most professions.
- Predictive analysis seems to be the holy grail for many managers.
- Prescriptive analysis uses the knowledge created in the previous phases to automatically run a business process and even decide on future courses of action.

Strategic Data Science

Content Reference : https://lucidmanager.org/data-science/strategic-data-science/ [Peter Prevos]

55





Data-Driven Decision Making Process

Content Reference : Veerasak Kritsanapraphan [Software Park Thailand]

56

- Curiosity
- Creativity
- Focus
- Attention to Detail

Characteristics of Data Scientist

57

- Finding rich data sources.
- Working with large volumes of data despite hardware, software, and bandwidth constraints.
- Cleaning the data and making sure that data is consistent.
- Melding multiple datasets together.
- Visualizing that data.

Data Scientist

Content Reference : Veerasak Kritsanapraphan [Software Park Thailand]

58

- Business User
- Project Sponsor
- Project Manager
- BI Analyst
- Database Administrator
- Data Engineer
- Data Scientist

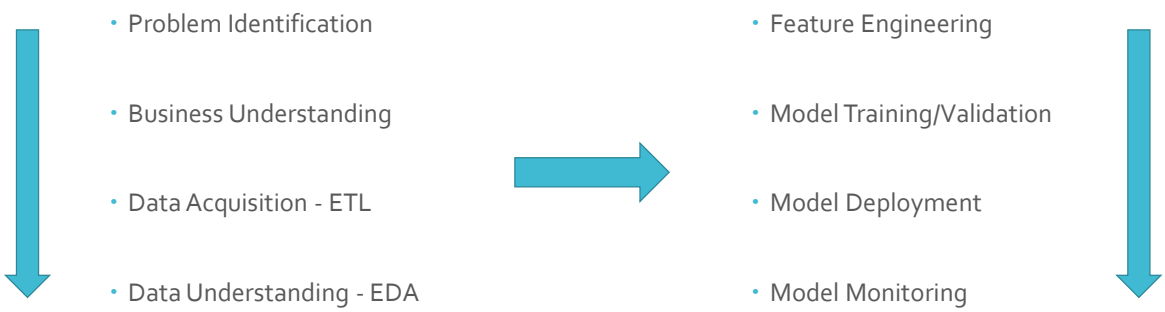
Data Science Team

59

- Transforming
- Creating
- As a Service
- Crowdsourcing

Approaches to Developing Data Science Capabilities

60



Data Science Life Cycles

61

- Frame the Business Problem
- Previous know outcome from reliable and different source and industry experts
- Figure out whether cost of Failure does create reputation damage
- Facilitate Experimentation with Automation
- Decide where you need to scale

Problem Identification

Content Reference : https://towardsdatascience.com [Sivakar Siva]

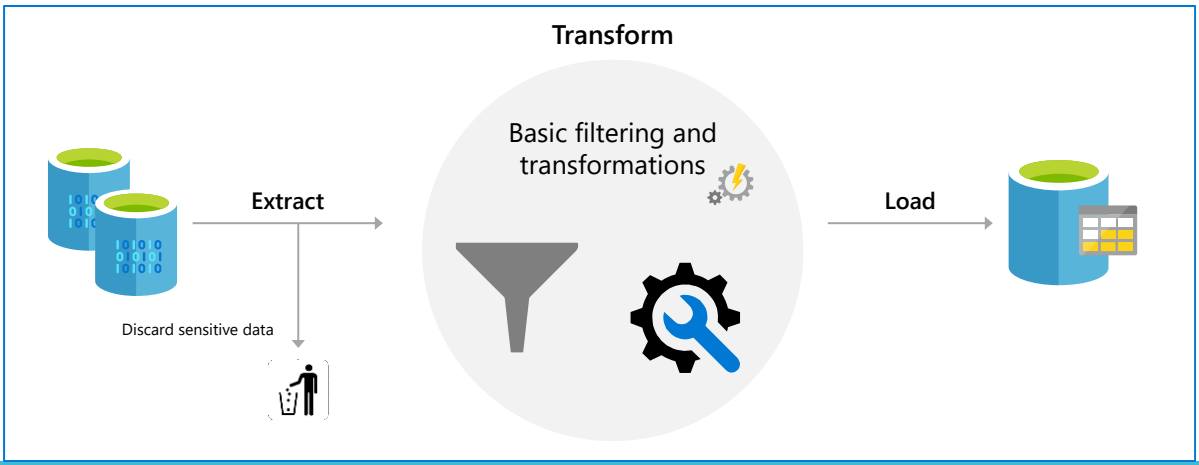
62

- Understand the Business Objective
- Understand scalability need
- Defining your success Criteria and success measuring Metrics KPI & SLA
- Identify the key business variables that the model needs to predict
- Discuss Integration of model with business process

Business Understanding

Content Reference : https://towardsdatascience.com [Sivakar Siva]

63

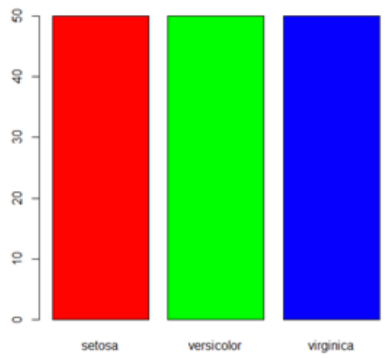


Data Acquisition - ETL

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

64

- Descriptive Statistics
  - Measures of Frequency
  - Measures of Central Tendency
  - Measures of Dispersion or Variation
  - Measures of Position
- Inferential Statistics
  - The estimation of the parameter(s)
  - Testing of statistical hypotheses.



Data Understanding - EDA

Content Reference : https://towardsdatascience.com [Sivakar Siva]

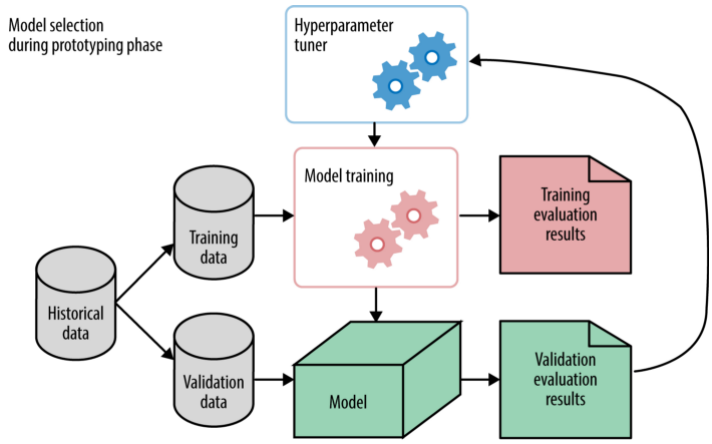
65

- Feature engineering is the process of using domain knowledge to extract features from raw data via data mining techniques.
- These **features** can be used to strengthen the performance of machine learning models.
- Feature engineering can be considered as applied machine learning itself.

Feature Engineering

Content Reference : https://towardsdatascience.com [Sivakar Siva]

66



Model Training/Validation

Content Reference : Oreilly by Alice Zheng

67

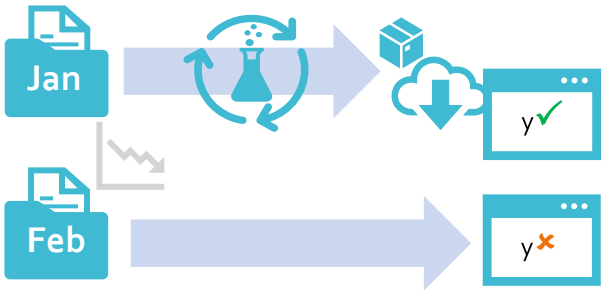
Deploy a Real-Time Pipeline

Publish a Batch Pipeline

Model Deployment

Content Reference : Microsoft Corporation

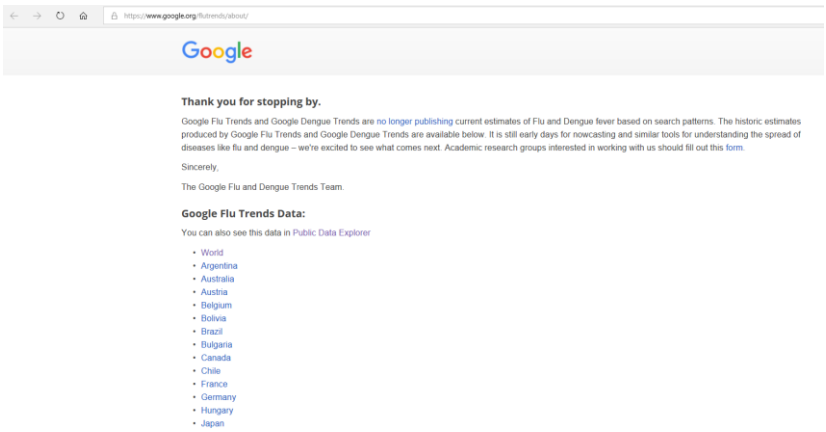
68



Model Monitoring

Content Reference : Microsoft Corporation

69



- **Google Flu Treads (Nowcasting Example)**
- https://www.google.org/flutrends/about/

Nowcasting vs Forecasting

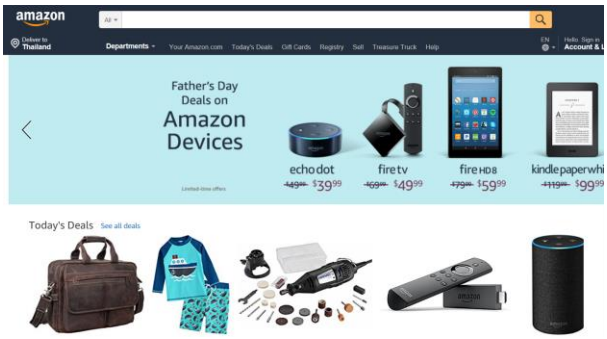
70

elections2012
(Forecasting Example)

Nowcasting vs Forecasting

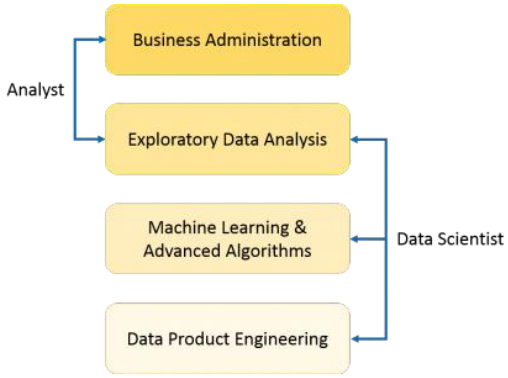
71

- **Transform Data Into …**
  - Valuable Insights
  - Data Productions
  - Interesting Stories



Summary of Data Science

72



- Predictive causal analytics
- Prescriptive analytics
- Machine learning for making predictions
- Machine learning for pattern discover

Summary of Data Science

Content Reference : https://www.edureka.co/blog/what-is-data-science/

73

03 - Big Data Fundamentals

74

- Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. But it's not the amount of data that's important. It's what organizations do with the data that matters.

- Big data can be analyzed for insights that lead to better decisions and strategic business moves.

Overview - Big Data

75

- **It's all happening online**
  - Click
  - Ad impression
  - Billing event
  - Fast Forward, pause,…
  - Server request
  - Transaction
  - Network message
  - Fault

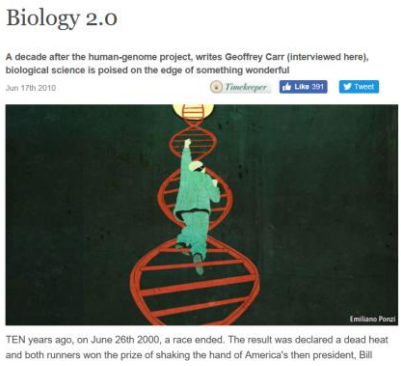
- **User Generate Content on Web & Mobile**

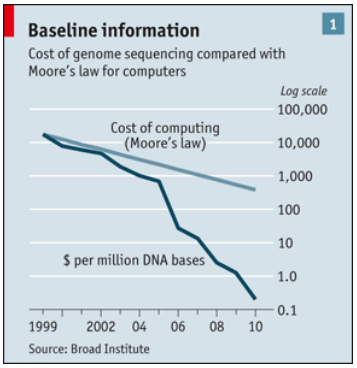


Where Dose Big Data Come From?

76

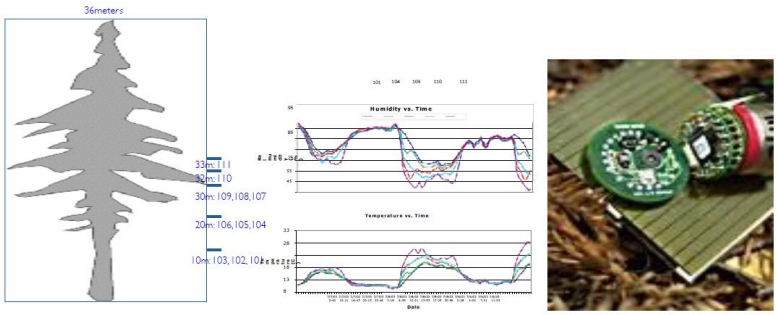
- Health and Scientific Computing



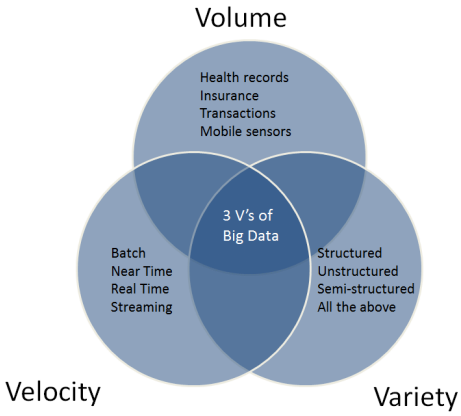


Where Dose Big Data Come From?

- Internet of Thing (IoT)



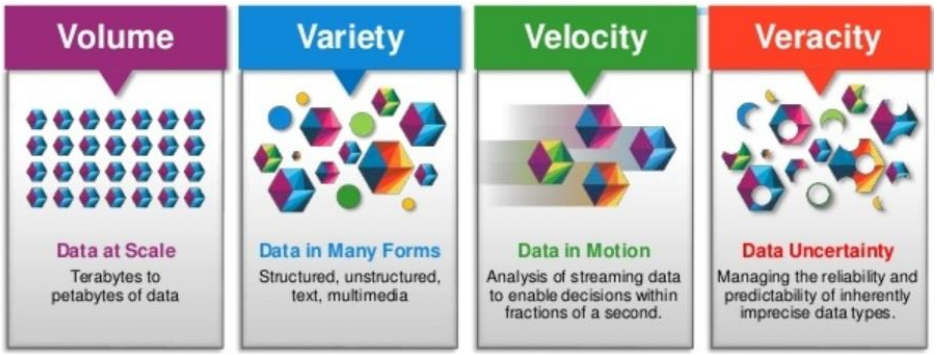
Where Dose Big Data Come From?



The 3 V's of Big Data Characteristics

Content Reference : Christopher O. Austin

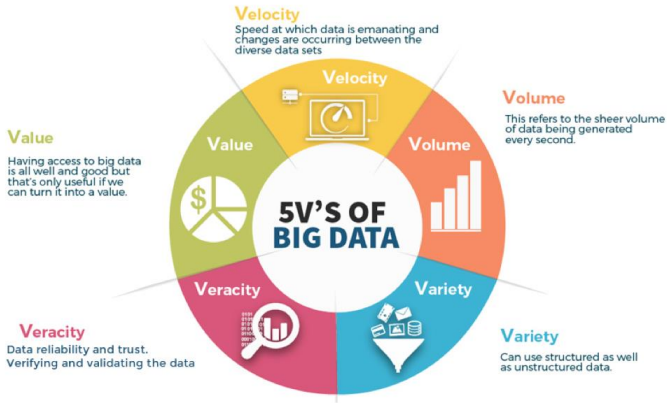
79



The 4V's of Big Data Characteristics (IBM)

Content Reference : https://www.sliceofbi.com/

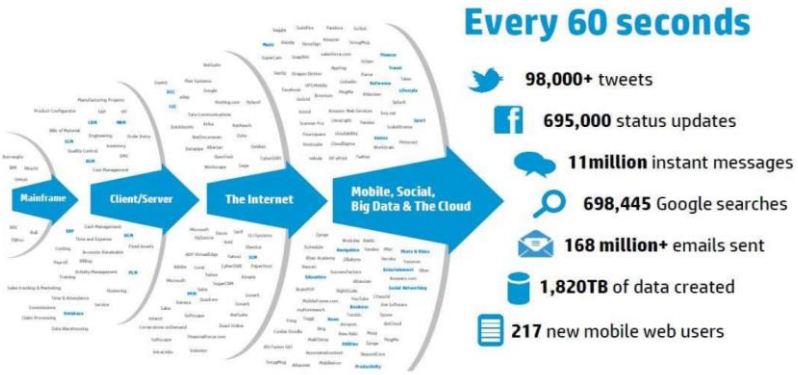
80



The 5V's of Big Data

Content Reference : https://www.techentice.com/the-data-veracity-big-data/

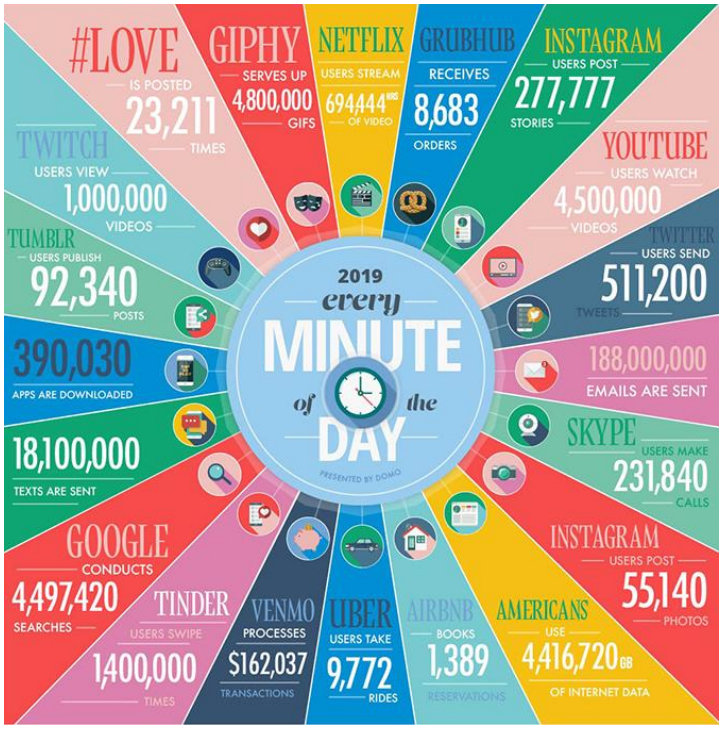
81



Big Data Characteristics

Content Reference : Bernard Marr's

82



83

Data Never Sleeps

Content Reference : https://web-assets.domo.com/blog/wp-content/uploads/2019/07/data-never-sleeps-7-896kb.jpg

https://www.youtube.com/watch?v=ZQPZ-_k_2Hg

84


- Hadoop – Big Data Platform
- MongoDB – Document Database [No-SQL]
- Cloud Computing



Big Data Technologies Overview

85

- Hadoop is an Apache open source framework written in java that allows **distributed processing of large datasets across clusters of computers** using simple programming models.
- A Hadoop framework application works in an environment that provides distributed storage and computation across clusters of computers.
- Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage.

Introduction to Hadoop

86

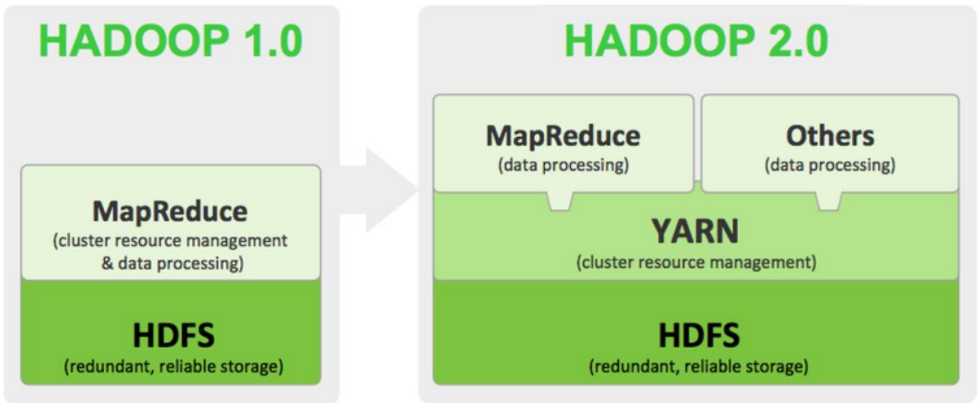
- When to use Hadoop
  - For processing really big data
  - For storing a diverse set of data
  - For parallel data processing

- When NOT to use Hadoop
  - For real-time data analysis
  - For a relational database system
  - For a general network file system
  - For non-parallel data processing

When (not) to use Hadoop

Content Reference : DigiTech [Business Insights Analytics Course]

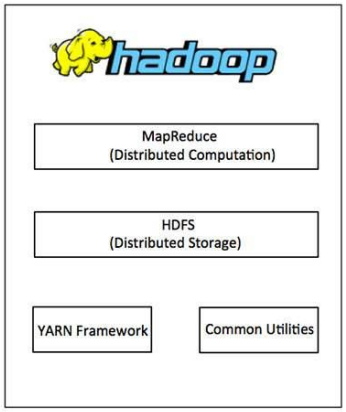
87



Hadoop History

Content Reference : DigiTech [Business Insights Analytics Course]

88



- **Hadoop Common**

These are Java libraries and utilities required by other Hadoop modules. These libraries provides filesystem and OS level abstractions and contains the necessary Java files and scripts required to start Hadoop.

- **Hadoop YARN**

This is a framework for job scheduling and cluster resource management.

- **Hadoop Distributed File System (HDFS)**

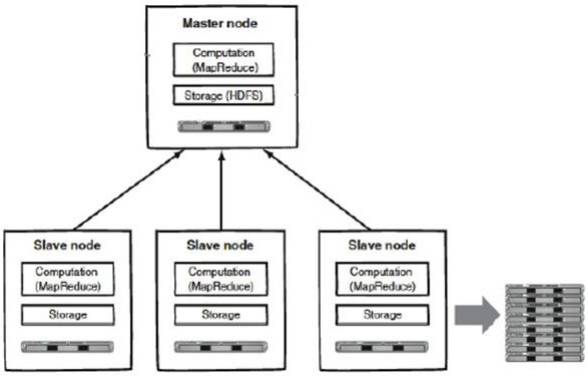
A distributed file system that provides high-throughput access to application data.

- **Hadoop MapReduce :**
  This is YARN-based system for parallel processing of large data sets.

Hadoop Architecture

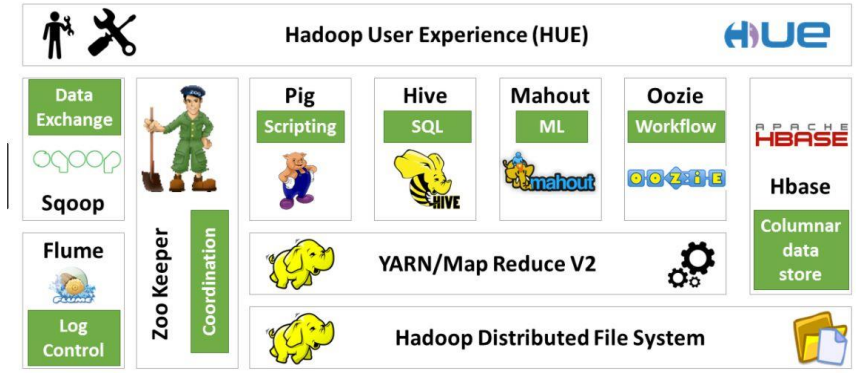
Content Reference : https://www.tutorialspoint.com/

89



Hadoop Architecture

90



Hadoop Ecosystem (Version 2)

Content Reference : https://www.ktexperts.com/hadoop-ecosystem [Neeraj Sharma]

91

- Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data summarization, query and analysis.
- Hive gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop.
- SQL-like queries (HiveQL)

Hadoop Technology - Apache Hive

92

- HiveQL Sample

```
1 DROP TABLE IF EXISTS docs;
2 CREATE TABLE docs (line STRING);
3 LOAD DATA INPATH 'input_file' OVERWRITE INTO TABLE docs;
4 CREATE TABLE word_counts AS
5 SELECT word, count(1) AS count FROM
6  (SELECT explode(split(line, '\s')) AS word FROM docs) temp
7 GROUP BY word
8 ORDER BY word;
```

Hadoop Technology - Apache Hive

93

- Apache Pig is a high-level platform for creating programs that run on Apache Hadoop.
- The language for this platform is called Pig Latin

```
1 input_lines = LOAD '/tmp/word.txt' AS (line:chararray);
2 words = FOREACH input_lines GENERATE FLATTEN(TOKENIZE(line)) AS word;
3 filtered_words = FILTER words BY word MATCHES '\\w+';
4 word_groups = GROUP filtered_words BY word;
5 word_count = FOREACH word_groups GENERATE COUNT(filtered_words) AS count, group AS word;
6 ordered_word_count = ORDER word_count BY count DESC;
7 STORE ordered_word_count INTO '/tmp/results.txt';
```

Hadoop Technology - Apache Pig

94

- HBase is an open-source, non-relational, distributed database modeled after Google's big table and is written in Java.
- HBase is not a direct replacement for a classic SQL database

Hadoop Technology - Apache HBase

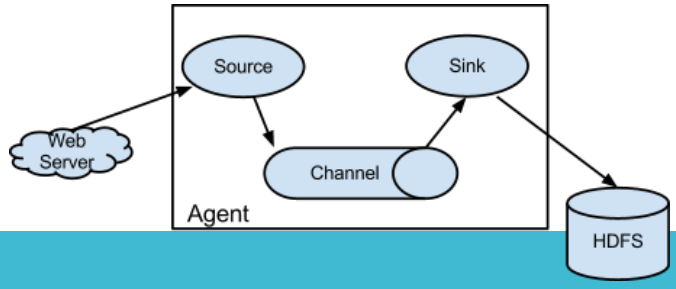
95

- Sqoop is a command-line interface application for transferring data between relational databases and Hadoop

Hadoop Technology - Apache Sqoop

96

- Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data.
- It has a simple and flexible architecture based on streaming data flows.



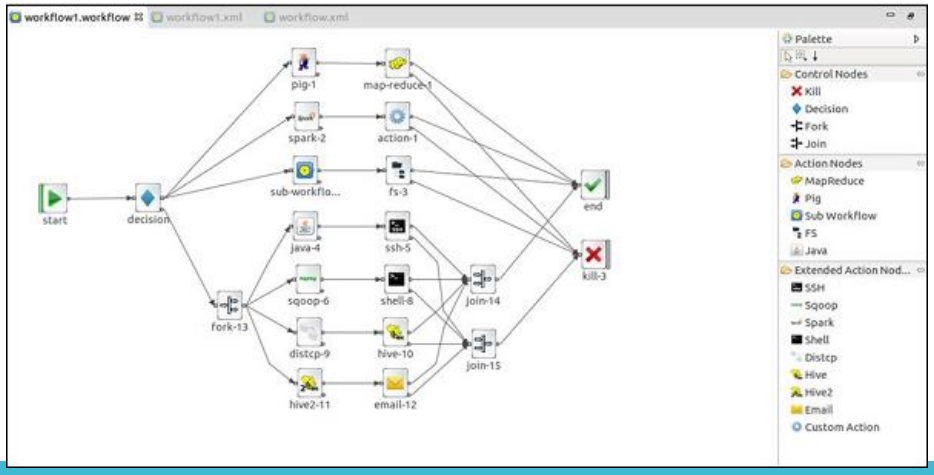
Hadoop Technology - Apache Flume

97

- Apache Oozie is a server-based workflow scheduling system to manage Hadoop jobs.
- Workflows in Oozie are defined as a collection of control flow and action nodes in a directed acyclic graph.
- Control flow nodes define the beginning and the end of a workflow (start, end, and failure nodes) as well as a mechanism to control the workflow execution path (decision, fork, and join nodes).

Hadoop Technology - Apache Oozie

98



Hadoop Technology - Apache Oozie

99

- Hue is an open source Analytics Workbench for browsing, querying and visualizing data.

Hadoop Technology - Apache Hue

100

- Apache Mahout is a project of the Apache Software Foundation to produce free implementations of distributed or otherwise scalable machine learning algorithms focused primarily in the areas of collaborative filtering, clustering and classification.
- Mahout also provides Java libraries for common maths operations (focused on linear algebra and statistics) and primitive Java collections.

Hadoop Technology - Apache Mahout

101

MongoDB is a database using document-oriented storage. Under this model, data is stored in documents and documents are combined into collections.

- Database
- Collection
- Document

Big Data Technology - MongoDB

102

RDBMS	MongoDB
Database	Database
Table	Collection
Tuple/Row	Document
column	Field
Table Join	Embedded Documents
Primary Key	Primary Key (Default key _id provided by mongodb itself)

Big Data Technology - MongoDB

103

- Schema less
- Structure of a single object is clear
- No complex joins
- Deep query-ability (document-based query language)
- Tuning
- Ease of scale-out

Advantages of MongoDB over RDBM

104

- Big Data
- Content Management and Delivery
- Mobile and Social Infrastructure
- User Data Management
- Data Hub

Where to Use MongoDB?

105

```
var MongoClient = require('mongodb').MongoClient;

//Create a database named "myDatabase":
var url = "mongodb://localhost:27017/myDatabase";

MongoClient.connect(url, function(err, db) {
  if (err) throw err;
  console.log("Database created!");
  db.close();
});
```

Create Database using NodeJS

106

```
var MongoClient = require('mongodb').MongoClient;
var url = "mongodb://localhost:27017/";

MongoClient.connect(url, function(err, db) {
  if (err) throw err;
  var dbo = db.db("myDatabase");

  //Create a collection name "customers":
  dbo.createCollection("customers",
   function(err, res) {
    if (err) throw err;
    console.log("Collection created!");
    db.close();
  });
});
```

Create Collection using NodeJS

107

```
var MongoClient = require('mongodb').MongoClient;
var url = "mongodb://localhost:27017/";

MongoClient.connect(url, function(err, db) {
  if (err) throw err;
  var dbo = db.db("myDatabase");

  var myobj = {
    name: "Bank of Thailand", address: "Tewet" };
  dbo.collection("customers").insertOne(myobj, function(err, res) {
    if (err) throw err;
    console.log("1 document inserted");
    db.close();
  });
});
```

Create Document using NodeJS

108

```
var MongoClient = require('mongodb').MongoClient;
var url = "mongodb://localhost:27017/";

MongoClient.connect(url, function(err, db) {
  if (err) throw err;
  var dbo = db.db("myDatabase");

    var myquery = { address: 'Bangkok' };
  dbo.collection("customers").deleteOne(myquery, function(err, obj)
  {
    if (err) throw err;
    console.log("1 document deleted");
    db.close();
  });
});
```

Delete Document using NodeJS

109

```
var MongoClient = require('mongodb').MongoClient;
var url = "mongodb://localhost:27017/";

MongoClient.connect(url, function(err, db) {
  if (err) throw err;
  var dbo = db.db("myDatabase");
  dbo.collection("customers")
  .drop(function(err, delOK) {
    if (err) throw err;
    if (delOK) console.log("Collection deleted");
    db.close();
  });
});
```

Drop Collection using NodeJS

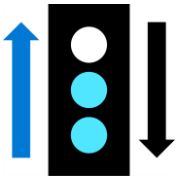
110

Cloud Computing

Compute

Storage

Cloud providers include Microsoft, Amazon, and Google



Networking

Analytics

Microsoft

111

Explore key cloud concepts

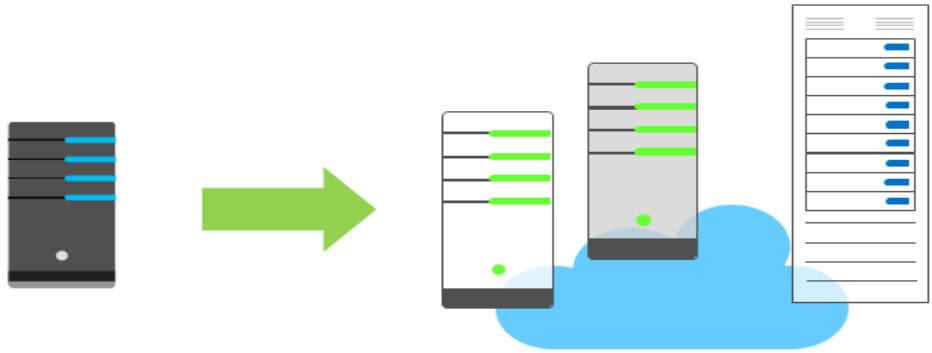
High availability	Fault tolerance
Scalability	Elasticity
Global reach	Customer latency capabilities
Agility	Predictive cost considerations
Disaster recovery	Security

Microsoft

112

Discuss economies of scale

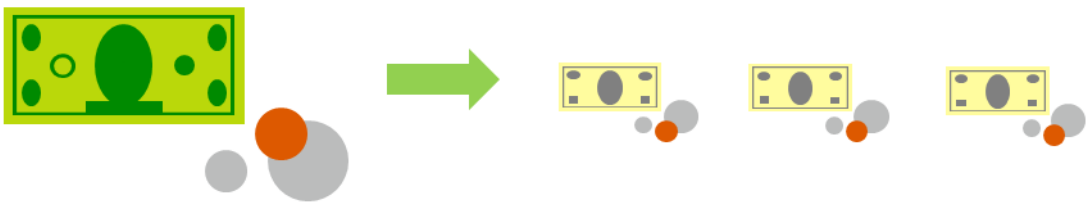
Economies of scale – Cloud providers can reduce costs and gain efficiency when operating at a large scale.



oft

113

Compare CapEx vs. OpEx



Capital Expenditure (CapEx)

- High upfront cost, value of investment reduces over time.

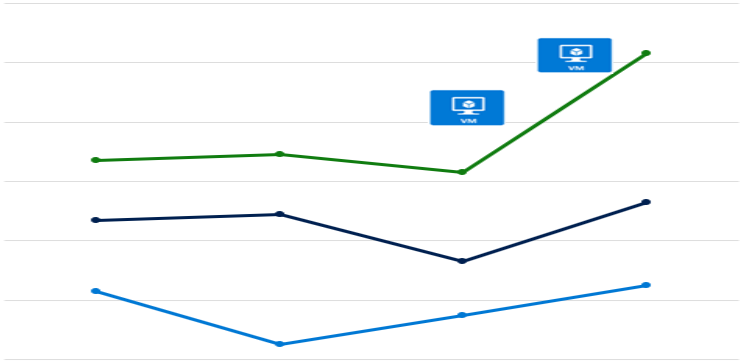
Operational Expenditure (OpEx)

- Spend on services or products as needed.
- No upfront cost, pay-as-you use.

Microsoft

114

Define consumption-based model

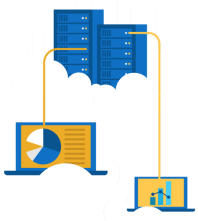


Consumption-based model = Pay only for the resources you use

Microsoft

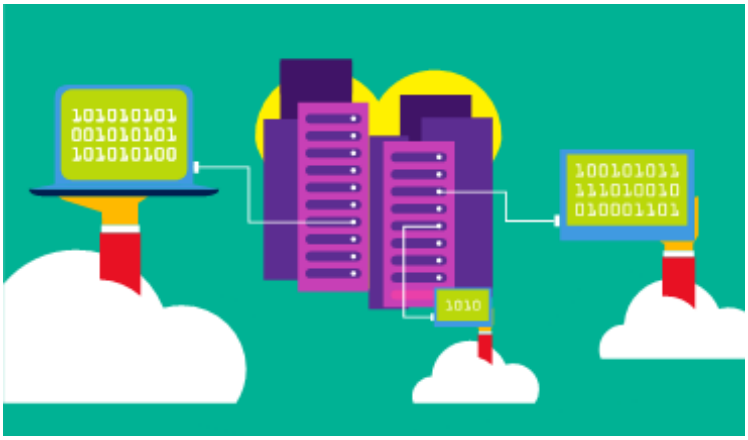
115

Distinguish types of cloud models



116

public cloud



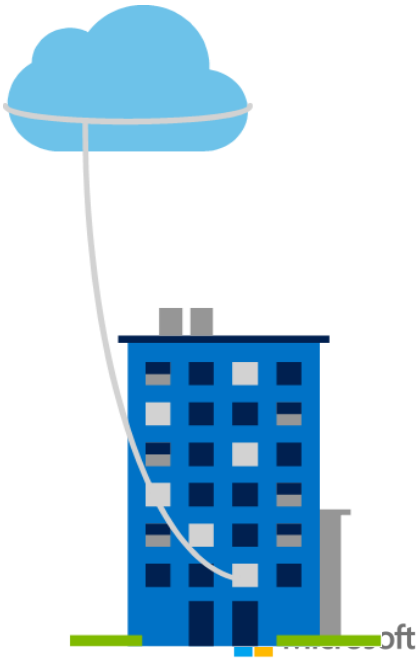
- Owned by cloud services or *hosting* provider.
- Provides resources and services to multiple organizations and users.
- Accessed via secure network connection (typically over the internet).

Microsoft

117

private cloud

- Organizations create a cloud environment in their datacenter.
- Organizations responsible for operating the services they provide.



118

hybrid cloud

Combines *Public* and *Private* clouds to allow applications to run in the most appropriate location.

Microsoft

119

Compare cloud models

Public cloud:	• No capital expenditures to scale up. • Applications can be quickly provisioned and deprovisioned. • Organizations pay only for what they use.
Private cloud:	• Organizations have complete control over resources. • Organizations have complete control over security.
Hybrid cloud:	• Most flexibility. • Organizations determine where to run their applications. • Organizations control security, compliance, or legal requirements.

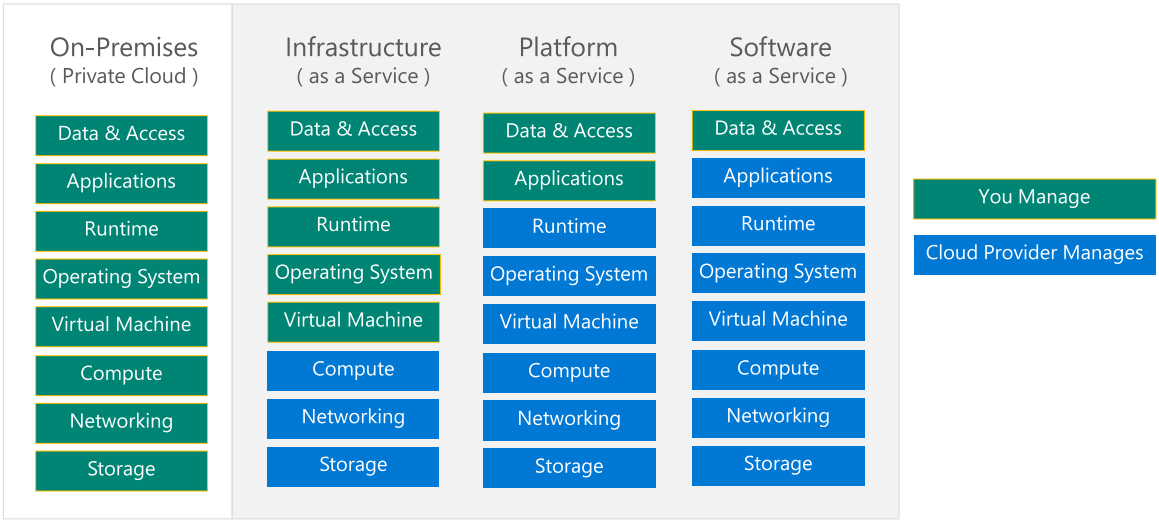
Microsoft

120

Explore types of cloud services

121

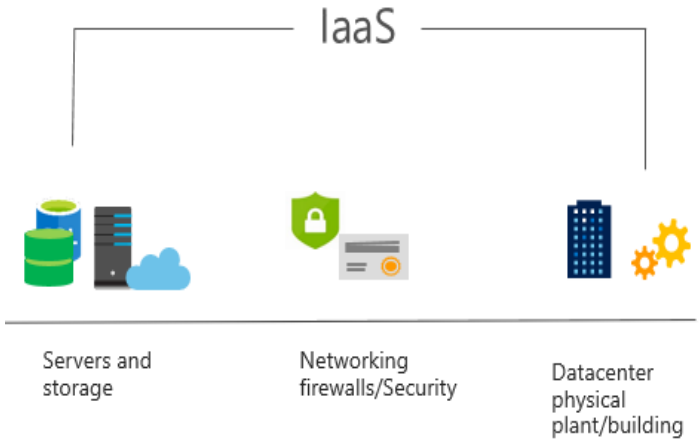
Discuss shared responsibility model



122

Define Infrastructure as a Service (IaaS)

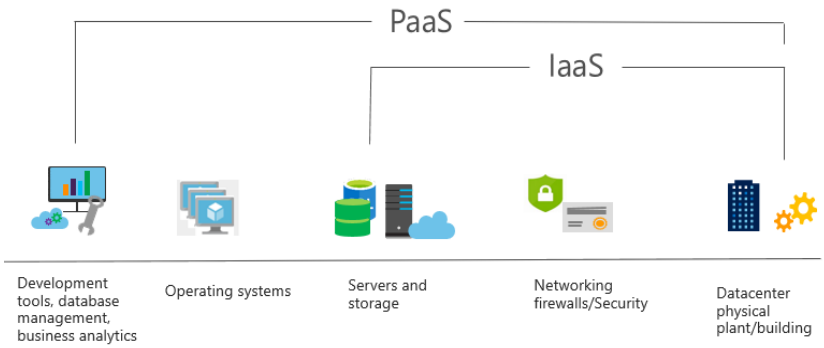
Build pay-as-you-go IT infrastructure by renting servers, virtual machines, storage, networks, and operating systems from a cloud provider.



Microsoft

123

Define Platform as a Service (PaaS)

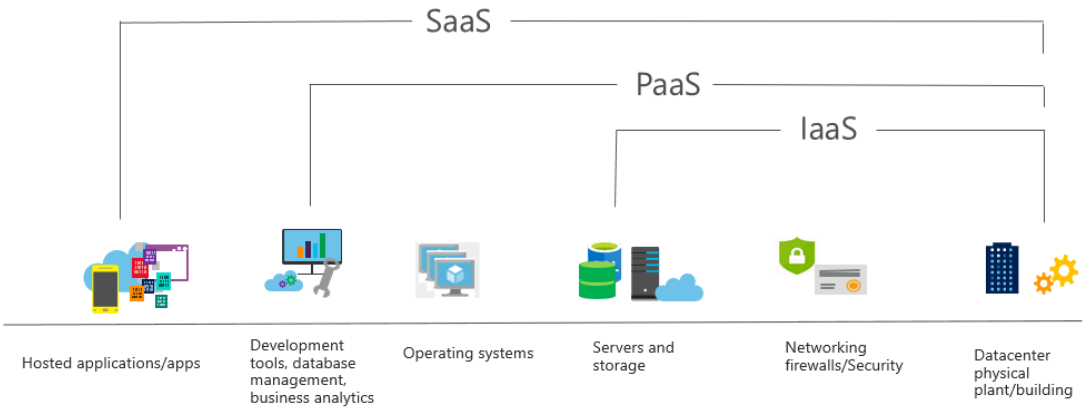


Provides environment for building, testing, and deploying software applications; without focusing on managing underlying infrastructure.

Microsoft

124

Define Software as a Service (SaaS)



Users connect to and use cloud-based apps over the internet: for example, Microsoft Office 365, email, and calendars.

Microsoft

125

Compare cloud services

IaaS	PaaS	SaaS
• The most flexible cloud service. • You configure and manage the hardware for your application.	• Focus on application development. • Platform management is handled by the cloud provider.	• Pay-as-you-go pricing model. • Users pay for the software they use on a subscription model.

Microsoft

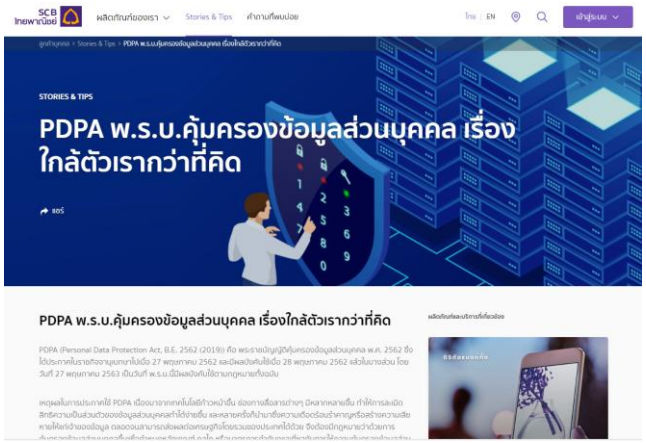
126

- Finance and Banking
- Telecommunication
- Retail
- Healthcare
- Agriculture
- Entertainment
- Real Estate

Big Data Use Case & Success Stories

127

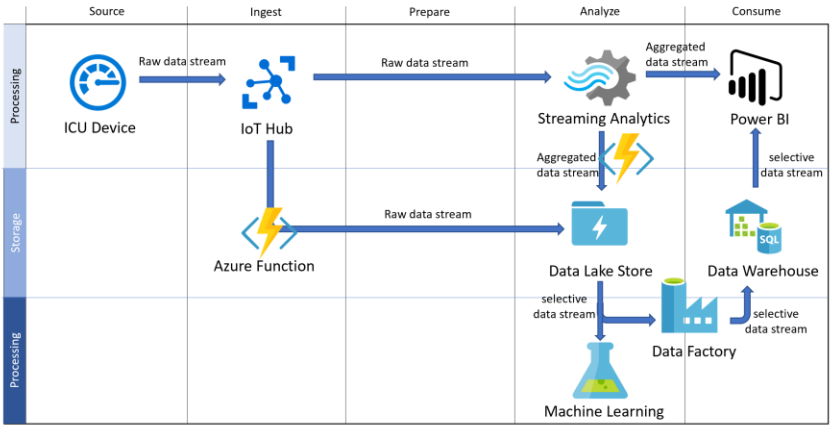
- Data privacy is responsibly collecting, using and storing data about people, in line with the expectations of those people, your customers, regulations and laws.
- Data ethics is doing the right thing with data, considering the human impact from all sides, and making decisions based on your brand values.



Privacy and Ethics of Big Data

Content Reference : https://looker.com/blog/big-data-ethics-privacy

128



Big Data Project – Case Example (Azure)

Content Reference : Microsoft Corporation

129

What to use for Data

Storage Account

- When you need a **low cost, high throughput** data store.
- When you need to store **No-SQL** data.
- When you **do not need to query** the data directly. **No ad hoc query** support.
- Suits the storage of archive or **relatively static data.**
- Suits acting as a **HDInsight Hadoop** data store.

Data Lake Store

- When you need a **low cost, high throughput** data store.
- **Unlimited storage** for **No-SQL** data
- When you **do not need to query** the data directly. **No ad hoc query** support.
- Suits the storage of archive or **relatively static data.**
- Suits acting as a **Databricks** , **HDInsight** and **IoT** data store.

Azure Databricks

- **Eases the deployment** of a Spark based cluster.
- Enables the **fastest processing** of Machine Learning solutions.
- **Enables collaboration** between data engineers and data scientists.
- Provides **tight enterprise security integration** with Azure Active Directory
- **Integration with other Azure Services** and **Power BI.**

Azure CosmosDB

- Provides **global distribution** for both structured and unstructured data stores.
- **Millisecond query response** time.
- **99.999% availability** of data.
- **Worldwide elastic scale** of both the storage and throughput
- **Multiple consistency levels** to control data integrity with concurrency

Azure SQL Database

- When you require a **relational** data store.
- When you need to manage **transactional workloads**
- When you need to manage a **high volume on inserts and reads**
- When you need a service that **requires high concurrency**
- When you require a solution that can scale **elastically**

130

What to use for Data

Azure Synapse Analytics

- When you require an integrated **relational** and **big data** store.
- When you need to manage **data warehouse** and **analytical workloads**
- When you need **low cost storage**.
- When you require the ability to **pause and restart the compute**.
- When you require a solution that can scale **elastically**

Azure Stream Analytics

- When you require a **fully managed event processing** engine.
- When you require **temporal analysis of streaming** data.
- Support for analyzing **IoT streaming** data.
- Support for analyzing application data through **Event Hubs.**
- Ease of use with a **Stream Analytics Query Language**.

Azure Data Factory

- When you want to **orchestrate the batch movement** of data.
- When you want to connect to **wide range of data platforms**.
- When you want to **transform or enrich** the data in movement.
- When you want to **integrate with SSIS packages**.
- Enables **verbose logging** of data processing activities.

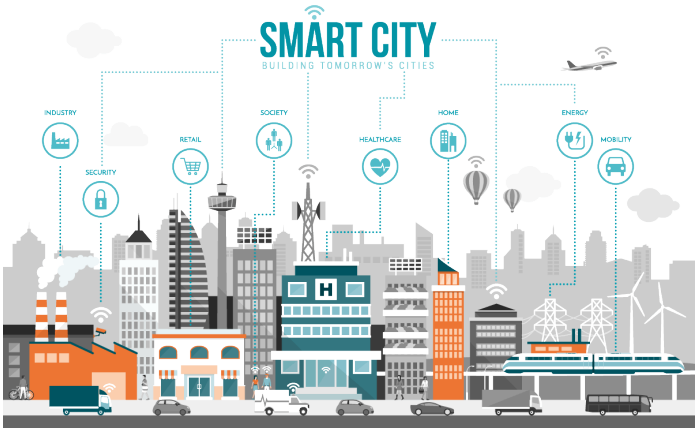
Azure HDInsight

- When you need a **low cost, high throughput** data store.
- When you need to store **No-SQL** data.
- Provides a Hadoop **Platform as a Service** approach
- Suits acting as a **Hadoop, Hbase, Storm or Kafka** data store.
- **Eases the deployment and management** of clusters.

Azure Data Catalog

- When you require **documentation** of your data stores.
- When you require a **multi user** approach to documentation.
- When you need to **annotate data sources** with descriptive metadata.
- A **fully managed cloud service** whose users can discover the data sources.
- When you require **a solution that can help business users** understand their data.

131



Smart City

Content Reference : https://www.arcweb.com/industries/smart-cities

132

Smart Farm

Content Reference : https://www.luda.farm/

133

04 - Big Data Sources

134

- Enterprise Data are functionally different application supporting the business operations in various departments.
- Enterprise Data Sources within an organization integrate and retrieve data for both internal applications and external communication.
- Enterprise Data Sources ensure trust and confidence in data assets.

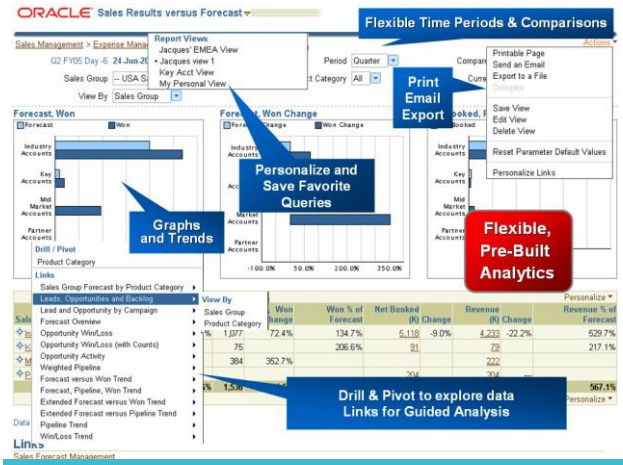
Introduction to Enterprise Data Source

135

- The vast majority of contemporary Enterprise Systems are based on relational database technology and, consequently, provide well-structured datasets that can be easily accessed.
- Enterprise Systems are Oracle, SQL Server, DB2, SAP, ….
- Use ODBC, JDBC, OLEDE, … For access to dataset.

Enterprise System

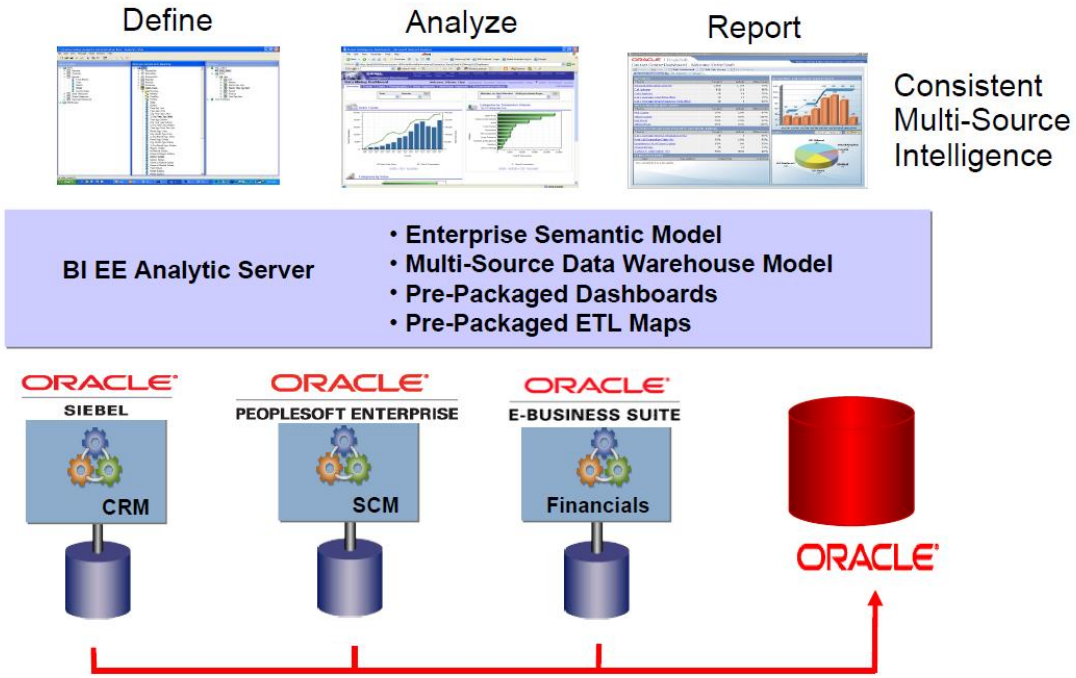
136



- Oracle provides a range of "Enterprise Application" including HR/Payroll, Finance, Customer Relationship Management (CRM).
- Oracle suite of application is Oracle E-Business Suite, PeopleSoft and JD Edwards.

Enterprise System - Oracle

137



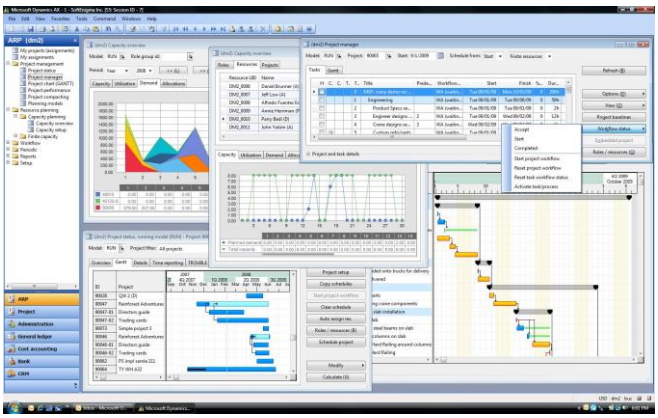
138

- SAP providers enterprise application that include modules such as HR/Payroll, Finance, and SAP Suite of offering also includes a number of vertical application.
- Extracting data from SAP is more difficult than other enterprise system since SAP users a proprietary language, ABAP, for data manipulation. Additionally, SAP database use tables which are different from standard database.
- In order to extract data from SAP we need to use additional tools.
  - SAP Connector from Oracle
  - SAP Business Connector
  - SAP Java Connector
  - SAP .NET Connector from SAP

Enterprise System - SAP

139

- Microsoft provides a set of enterprise application through its Dynamic suite. The suite includes CRM and ERP modules and is targeted at small and medium enterprises.
- The Microsoft Dynamics Suite is based on relational database technology (SQL Server)



Enterprise System - Microsoft

140

- Defined in many different ways, but not rigorously.
  - A decision support database that is maintained separately from the organization's operational database
  - Support information processing by providing a solid platform of consolidated, historical data for analysis.
- "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."—W. H. Inmon
- Data warehousing: The process of constructing and using data warehouses

Data Warehouse

Content Reference : Jiawei Han, Micheline Kamber, and Jian Pei University of Illinois at Urbana-Champaign & Simon Fraser University

141

- Organized around major subjects, such as customer, product, sales
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing
- Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process

Data Warehouse : Subject-Oriented

Content Reference : Jiawei Han, Micheline Kamber, and Jian Pei University of Illinois at Urbana-Champaign & Simon Fraser University

142

- Constructed by integrating multiple, heterogeneous data sources
  - relational databases, flat files, on-line transaction records
- Data cleaning and data integration techniques are applied.
  - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources E.g., Hotel price: currency, tax, breakfast covered, etc.
  - When data is moved to the warehouse, it is converted.

Data Warehouse : Integrated

Content Reference : Jiawei Han, Micheline Kamber, and Jian Pei University of Illinois at Urbana-Champaign & Simon Fraser University

143

- The time horizon for the data warehouse is significantly longer than that of operational systems
  - Operational database: current value data
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)
- Every key structure in the data warehouse
  - Contains an element of time, explicitly or implicitly
  - But the key of operational data may or may not contain "time element"

Data Warehouse : Time Variant

Content Reference : Jiawei Han, Micheline Kamber, and Jian Pei University of Illinois at Urbana-Champaign & Simon Fraser University

144

- A physically separate store of data transformed from the operational environment
- Operational update of data does not occur in the data warehouse environment
  - Does not require transaction processing, recovery, and concurrency control mechanisms
  - Requires only two operations in data accessing: initial loading of data and access of data

Data Warehouse : Nonvolatile

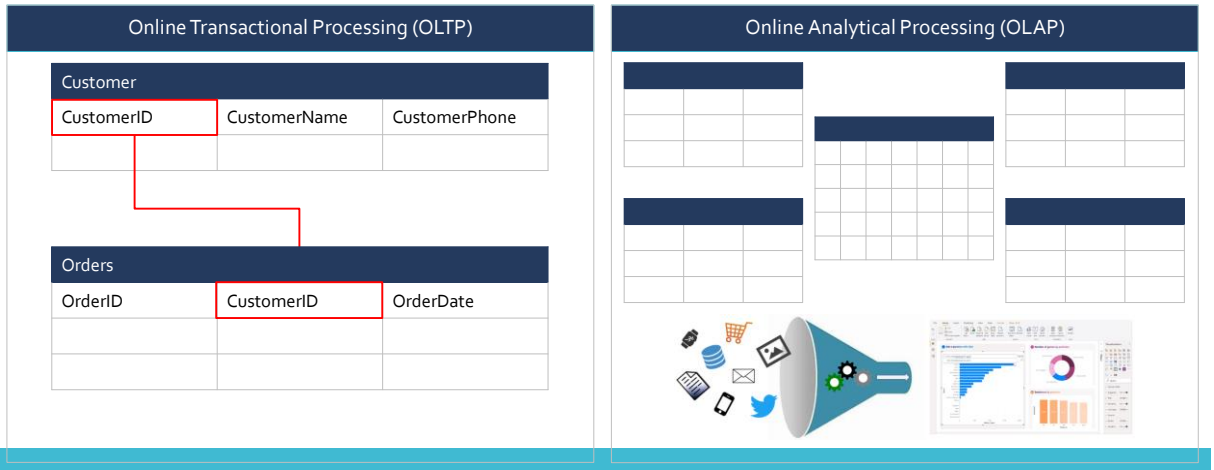
Content Reference : Jiawei Han, Micheline Kamber, and Jian Pei University of Illinois at Urbana-Champaign & Simon Fraser University

145

	OLTP	OLAP
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

Content Reference : Jiawei Han, Micheline Kamber, and Jian Pei University of Illinois at Urbana-Champaign & Simon Fraser University

146



Transactional vs analytical data stores

Content Reference : Microsoft Corporation [DP-900 : Microsoft Azure Data Fundamentals]

147

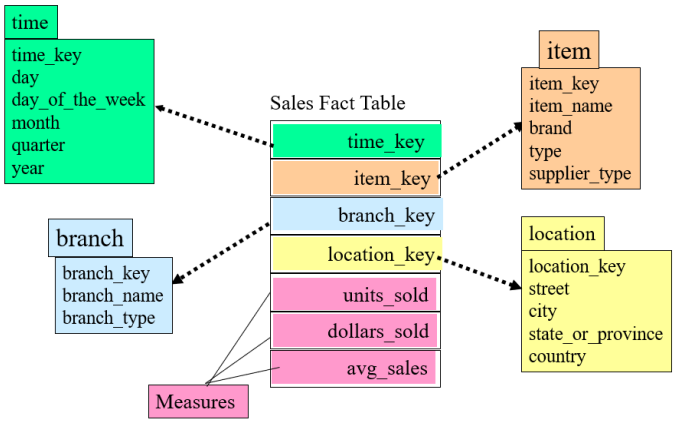
Modeling data warehouses: dimensions & measures

- **Star schema**: A fact table in the middle connected to a set of dimension tables
- **Snowflake schema**: A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake
- **Fact constellations**: Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation

Conceptual Modeling of Data Warehouses

Content Reference : Jiawei Han, Micheline Kamber, and Jian Pei University of Illinois at Urbana-Champaign & Simon Fraser University

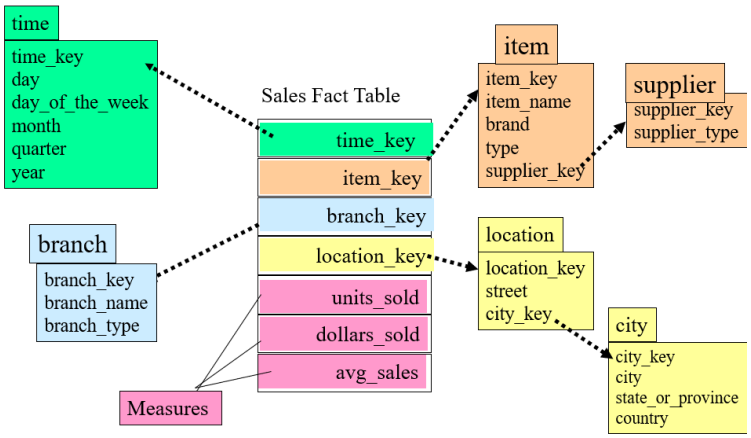
148



Example of Star Schema

Content Reference : Jiawei Han, Micheline Kamber, and Jian Pei University of Illinois at Urbana-Champaign & Simon Fraser University

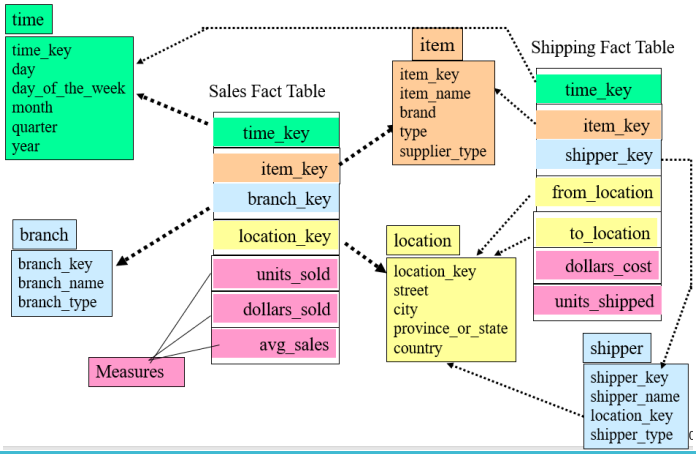
149



Example of Snowflake Schema

Content Reference : Jiawei Han, Micheline Kamber, and Jian Pei University of Illinois at Urbana-Champaign & Simon Fraser University

150

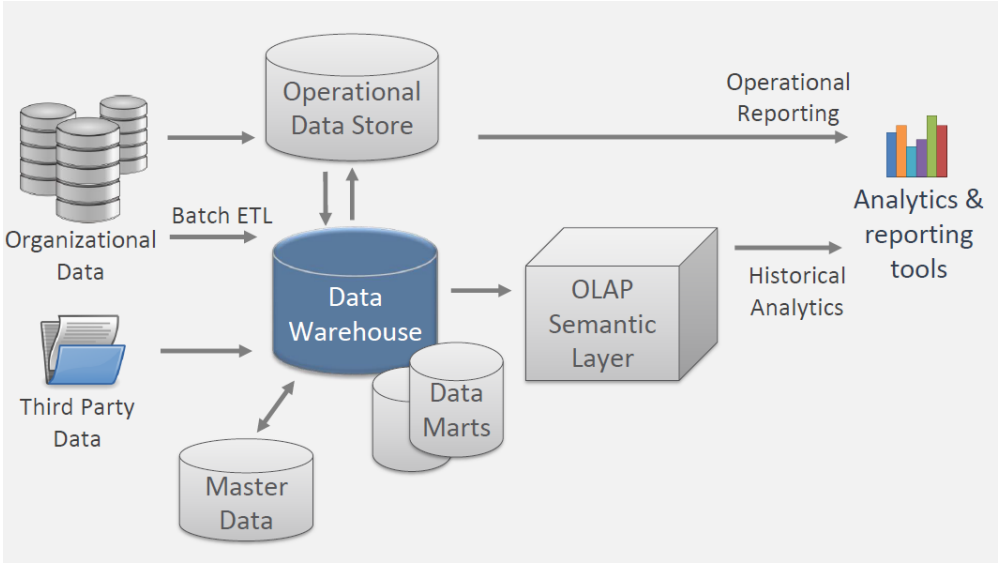


Example of Fact Constellation

Content Reference : Jiawei Han, Micheline Kamber, and Jian Pei University of Illinois at Urbana-Champaign & Simon Fraser University

151

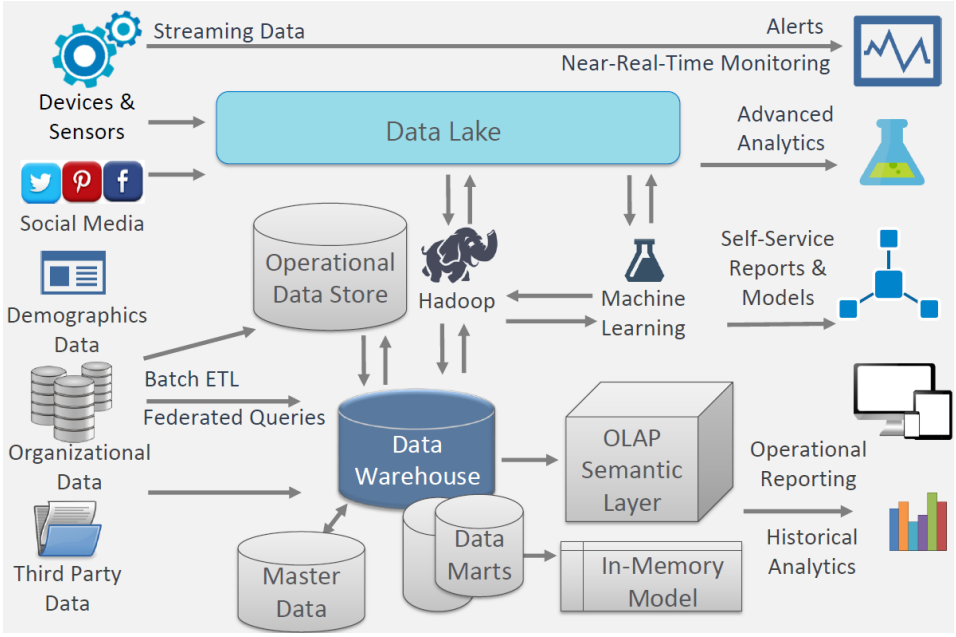
Traditional Data Warehousing



Content Reference : Melissa Coates [Blog: sqlchick.com/Twitter: @sqlchick]

152

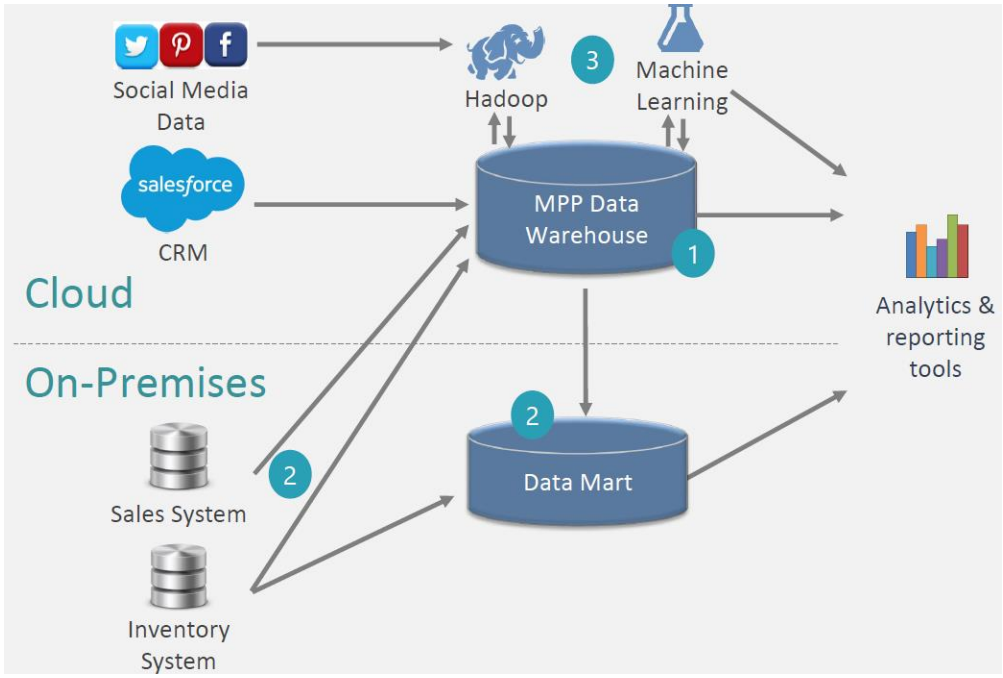
Modern Data Warehouses



Content Reference : Melissa Coates [Blog: sqlchick.com/Twitter: @sqlchick]

153

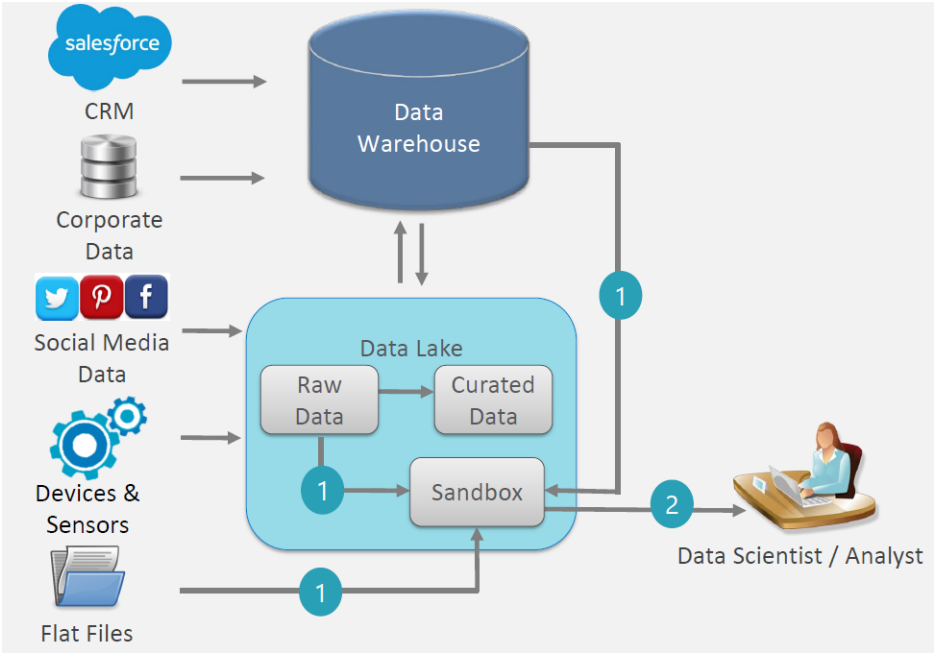
Hybrid Architecture



Content Reference : Melissa Coates [Blog: sqlchick.com/Twitter: @sqlchick]

154

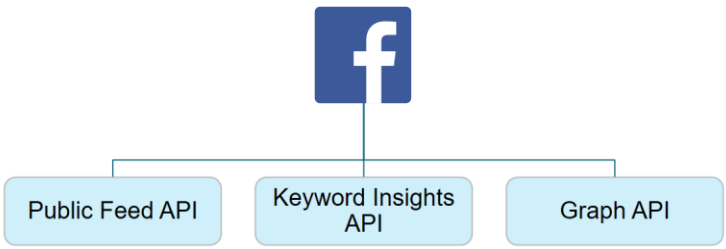
Sandbox Solutions



Content Reference : Melissa Coates [Blog: sqlchick.com/Twitter: @sqlchick]

155

- Facebook provides several ways to access its data streams. The access is done through separate Application Programming Interface (APIs)
- Some APIs are available only to certain organizations and currently cannot be accessed by the general public.

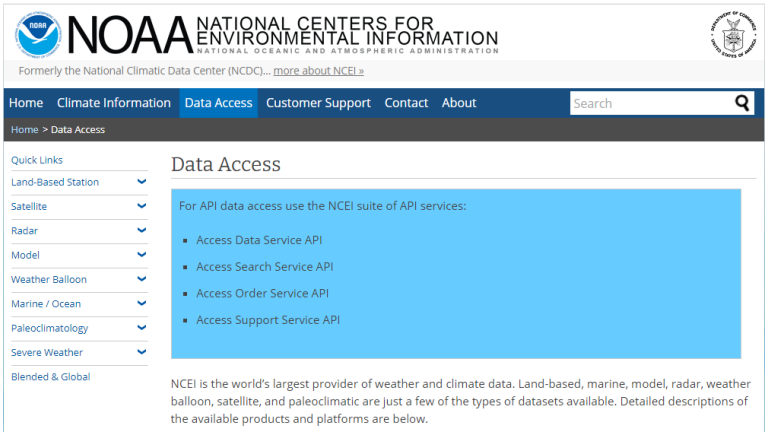


Social Media Data Sources - Facebook

156

Social Media Data Sources - Facebook

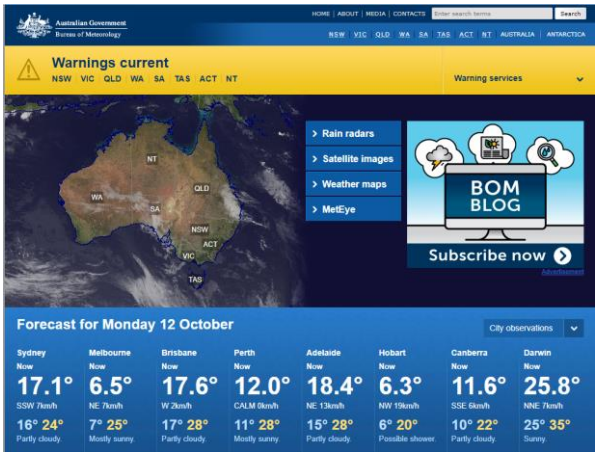
157



Public Data Sources - Weather

Content Reference : https://www.ncdc.noaa.gov/data-access

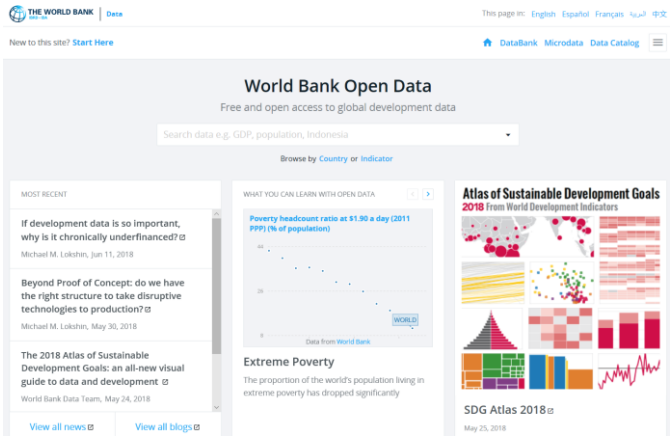
158



Public Data Sources - Weather

Content Reference : http://www.bom.gov.au/

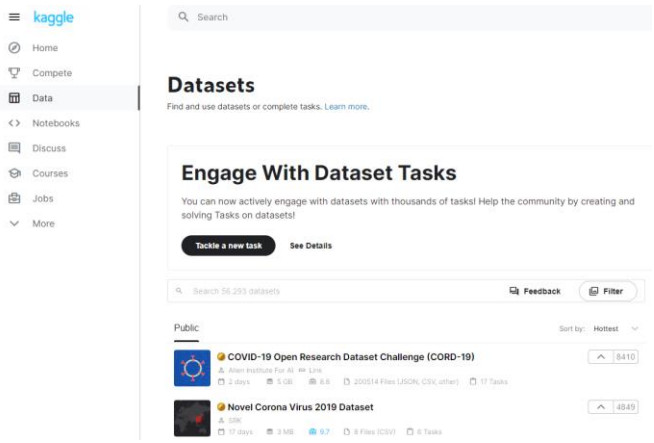
159



Public Data Sources – Economics & Finance

Content Reference : https://data.worldbank.org/

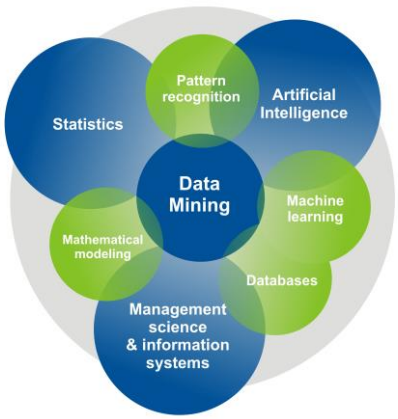
160



kaggle

Content Reference : https://www.kaggle.com/datasets

161



- Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.

Introduction to Data Mining

Content Reference : https://olucampbellcom.wordpress.com/

162

- The explosive growth in data collection
  - The storing of data in data warehouses.
  - The availability of increased access to data from Web navigation and intranet.

*** **We have to find a more effective way to use these data in decision support process than just using traditional query languages** ***

Why Data Mining?

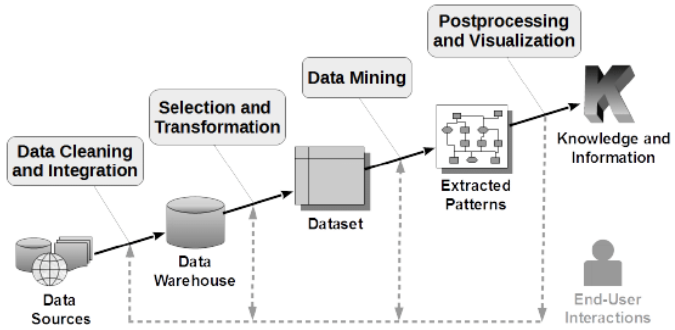
163

- Knowledge discovery in databases
- Knowledge extraction
- Data archeology
- Data exploration
- Data pattern processing
- Data dredging

Data Mining - Objectives

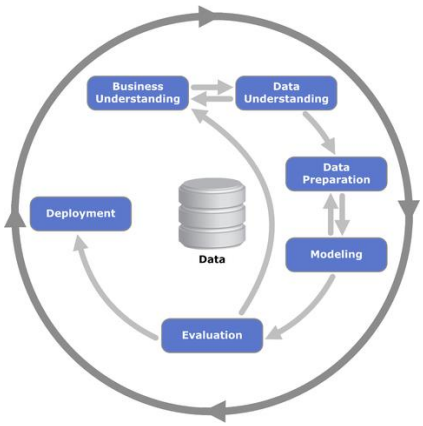
164

- Data Cleaning
- Data Integration
- Data Selection
- Data Transformation
- Data Mining
- Pattern Evaluation
- Knowledge Representation



Data Mining Process

165



Cross-Industry process for data mining

Data Mining Process - CRISP-DM

Content Reference : https://www.stitchdata.com/resources/what-is-data-mining/

166

- Prediction
  - how certain attributes within the data will behave in the future.
- Identification
  - identify the existence of an item, an event, an activity.
- Classification
  - partition the data into categories.
- Optimization
  - optimize the use of limited resources.

Data Mining - Goals

167

- **Association**
  - Providing the rules correlate the presence of a set of items with another set of item.
- **Classification**
  - Classification is the process of learning a model that describes different classes of data, the classes are predetermined.
  - The model that is produced is usually in the form of a decision tree or a set of rules.
- **Clustering**
  - The previous data mining task of classification deals with partitioning data based on a pre-classified training sample.

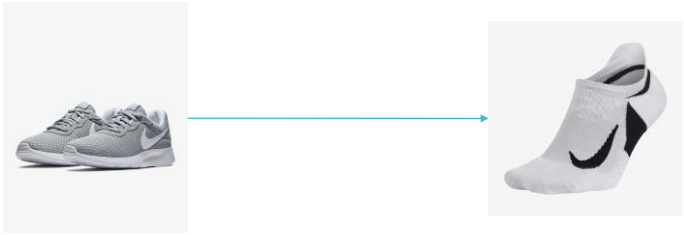
Basic Types of Data Mining

168

Market-basket model.

- Look for combinations of products.
- Put the SHOES near the SOCKS so that if a customer buys one they will buy the other.

Transactions: is the fact the person buys some items in the itemset at supermarket.



Association

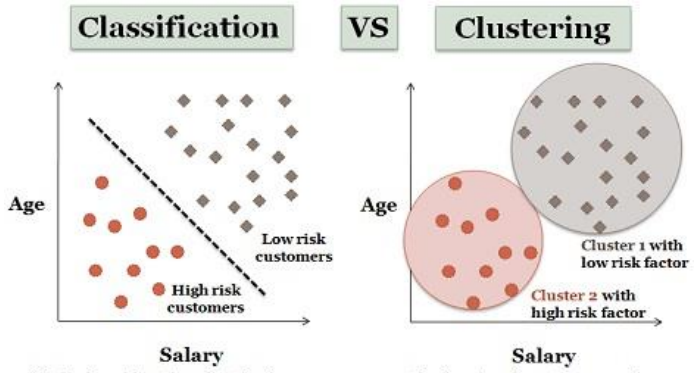
169

case ID | predictors | target

CUST_ID	CUST_GENDER	EDUCATION	OCCUPATION	AGE	AFFINITY_CARD
101501	F	Masters	Prof.	41	0
101502	M	Bach.	Sales	27	0
101503	F	HS-grad	Cleric.	20	0
101504	M	Bach.	Exec.	45	1
101505	M	Masters	Sales	34	1
101506	M	HS-grad	Other	38	0
101507	M	< Bach.	Sales	28	0
101508	M	HS-grad	Sales	19	0
101509	M	Bach.	Other	52	0
101510	M	Bach.	Sales	27	1

Classification

170



Risk classification for the loan payees on the basis of customer salary

Classification vs Clustering

Content Reference : https://www.educba.com/data-mining-methods/

171

- Weka is data mining software that uses a collection of machine learning algorithms.
- These algorithms can be applied directly to the data or called from the Java code.
- Weka is a collection of tools for:
  - Regression
  - Clustering
  - Association
  - Data pre-processing
  - Classification
  - Visualization

Data Mining Tools - Weka

Content Reference : https://www.cs.waikato.ac.nz/ml/weka/

172

- KNIME is a free and open-source data analytics, reporting and integration platform.
- KNIME integrates various components for machine learning and data mining through its modular data pipelining concept.

Data Mining Tools - KNIME

Content Reference : https://www.knime.com

173

Business Applications

- Predict behavior
- Deliver personalized services
- Measure profitability

Challenges

- Noisy data
- Scalability
- Incomplete data

Data Mining Summary

Content Reference : https://www.stitchdata.com/resources/what-is-data-mining/

174

Best Practices

- Preserve data
- Have a good idea of the insights you seek
- Strive for data quality
- Recognize outliers

Data Mining Summary

Content Reference : https://www.stitchdata.com/resources/what-is-data-mining/

175

05 - Introduction to Data Analytics

176

Data Science - Analytic Type

- Analysis
- Business Analytics

Data Science - Build Model

- Build
- AI & Machine Learning

Two Type of Data Science

Content Reference : Dr. Thanachart Ritbumroong [NIDA]

177

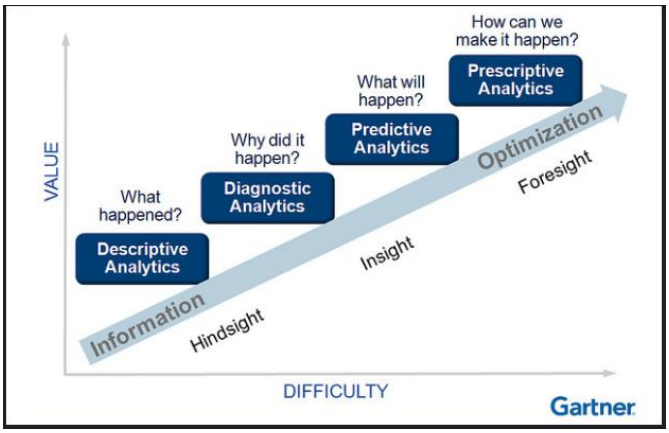
Microsoft Power Platform

Overview of Data Analysis

Data Analysis is telling a story with data.

Five categories of analytics:
- Descriptive
- Diagnostic
- Predictive
- Prescriptive
- Cognitive

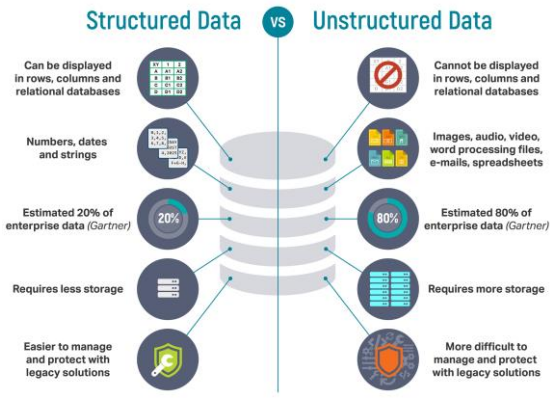
178

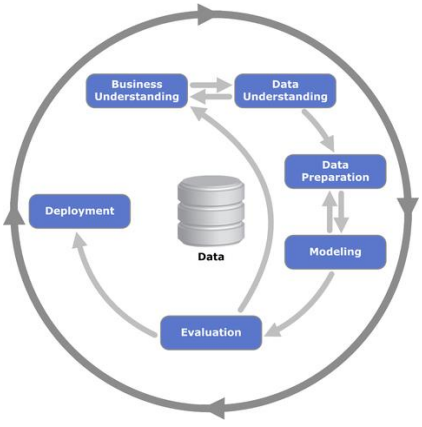


Data Analytics Maturity

Content Reference : Gartner

179





Understanding the Nature of the Data and The Data Analysis Process

Content Reference : https://www.igneous.io/blog/structured-data-vs-unstructured-data

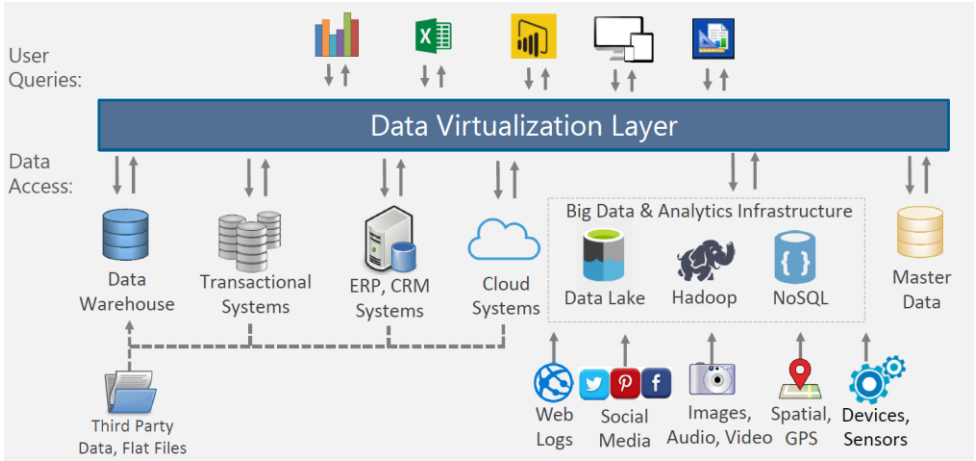
180

- Customer Lifetime Value
- Customer Segmentation
- Up and Cross Selling
- Next Best Action
- Propensity to buy
- Chun Prevention
- Fraud Detection
- Risk Management
- Demand Forecast
- Price Optimization
- Quality Assurance
- Predictive Maintenance

Data Analytics in Business

Content Reference : https://www.stitchdata.com/resources/what-is-data-mining/

181



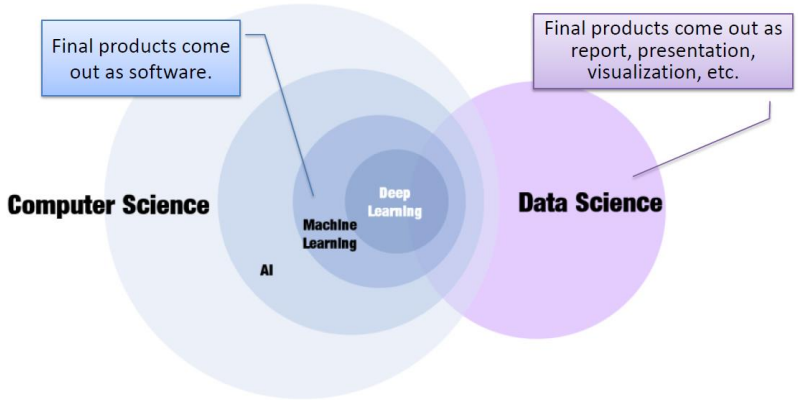
Data Virtualization

Content Reference : Melissa Coates [Blog: sqlchick.com/Twitter: @sqlchick]

182

06 - Understanding Machine Learning

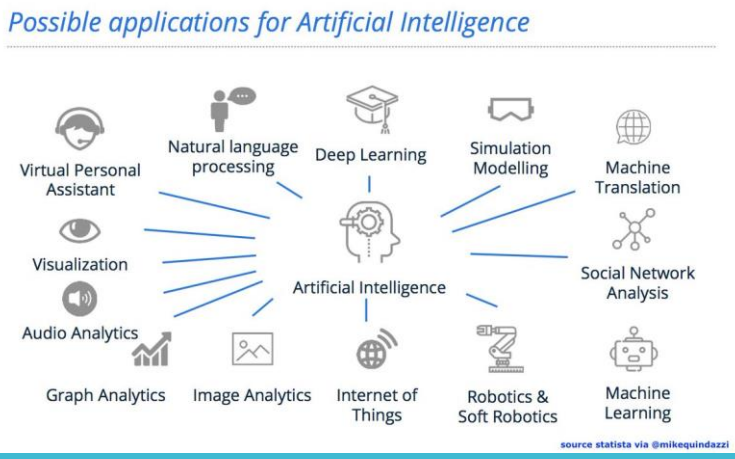
183



Terminology

Content Reference : https://www.coursera.org/learn/ai-for-everyone

184



Artificial Intelligence (AI)

Content Reference : statista via @mikequindazzi

185

Software that imitates human capabilities

- Making decisions based on data and past experience
- Recognizing abnormal events
- Interpreting visual input
- Understanding written and spoken language
- Engaging in dialogs and conversations

What is Artificial Intelligence?

Content Reference : Microsoft Corporation

186

	Machine Learning	Predictive models based on data and statistics – the foundation for AI
	Anomaly Detection	Systems that detect unusual patterns or events, enabling pre-emptive action
	Computer Vision	Applications that interpret visual input from cameras, images, or videos
	Natural Language Processing	Applications that can interpret written or spoken language
	Conversational AI	AI agents, (or *bots*), that can engage in dialogs with human users

Common Artificial Intelligence Workloads

Content Reference : Microsoft Corporation

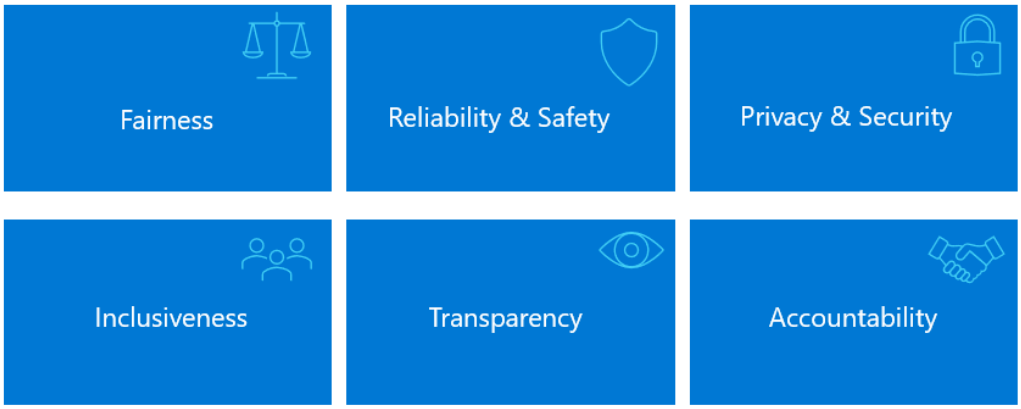
187

Challenge or Risk	Example
Bias can affect results	A loan-approval model discriminates by gender due to bias in the data with which it was trained
Errors may cause harm	An autonomous vehicle experiences a system failure and causes a collision
Data could be exposed	A medical diagnostic bot is trained using sensitive patient data, which is stored insecurely
Solutions may not work for everyone	A predictive app provides no audio output for visually impaired users
Users must trust a complex system	An AI-based financial tool makes investment recommendations - what are they based on?
Who's liable for AI-driven decisions?	An innocent person is convicted of a crime based on evidence from facial recognition – who's responsible?

Challenges and Risks with AI

Content Reference : Microsoft Corporation

188

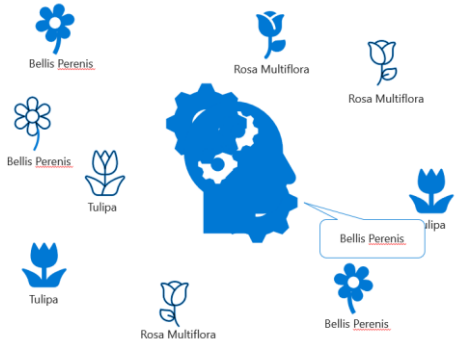


Principles of Responsible AI

Content Reference : Microsoft Corporation

189

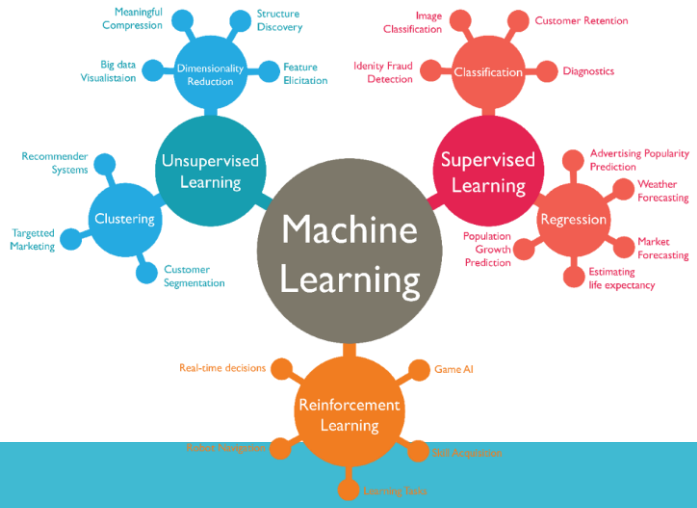
Creating predictive models by finding relationships in data



What is Machine Learning

Content Reference : Microsoft Corporation

190



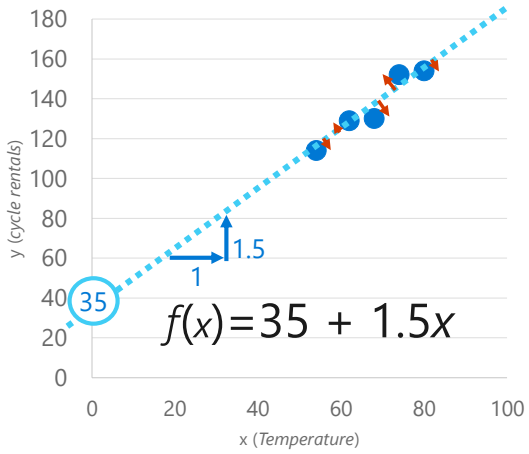
Type of Machine Learning

Content Reference : https://www.7wdata.be/visualization/types-of-machine-learning-algorithms-2/

191

Regression

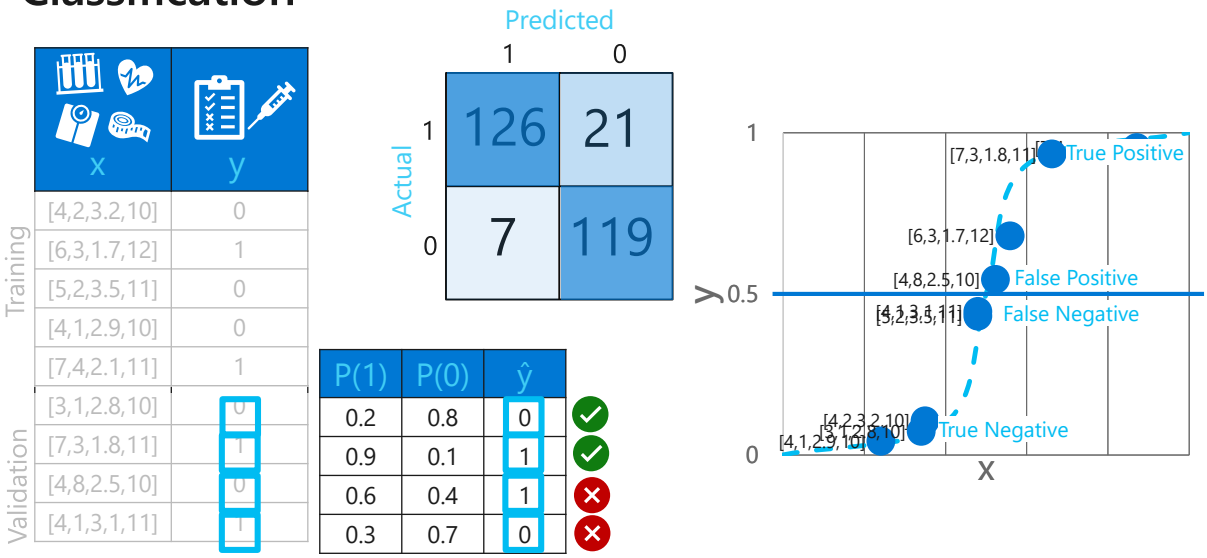
	x	y	$f(x)$ $\hat{y}$
Training	56	115	
	61	126	
	67	137	
	72	140	
	76	152	
	82	156	
Validation	54	114	116
	62	129	128
	68	130	137
	74	152	146
	80	154	155



Content Reference : Microsoft Corporation

192

Classification

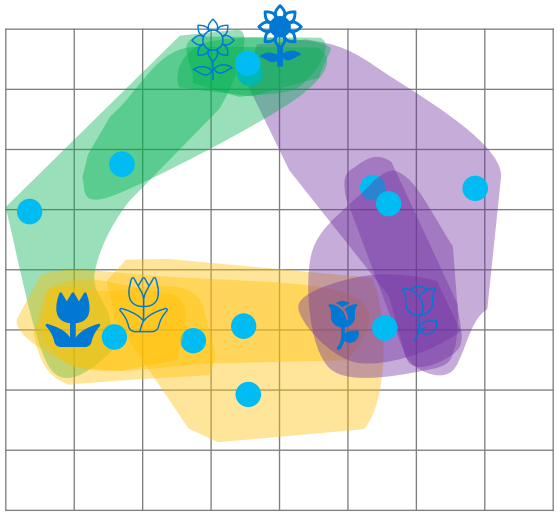


Content Reference : Microsoft Corporation

193

Clustering

	6	3
	5	3
	2	3
	1	3
	3	8
	4	8



Content Reference : Microsoft Corporation

194

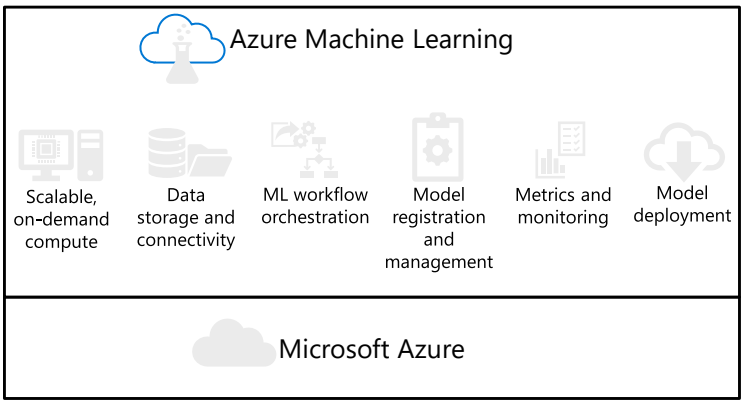
Microsoft

Azure Machine Learning No-Code with Designer

195

What is Azure Machine Learning?

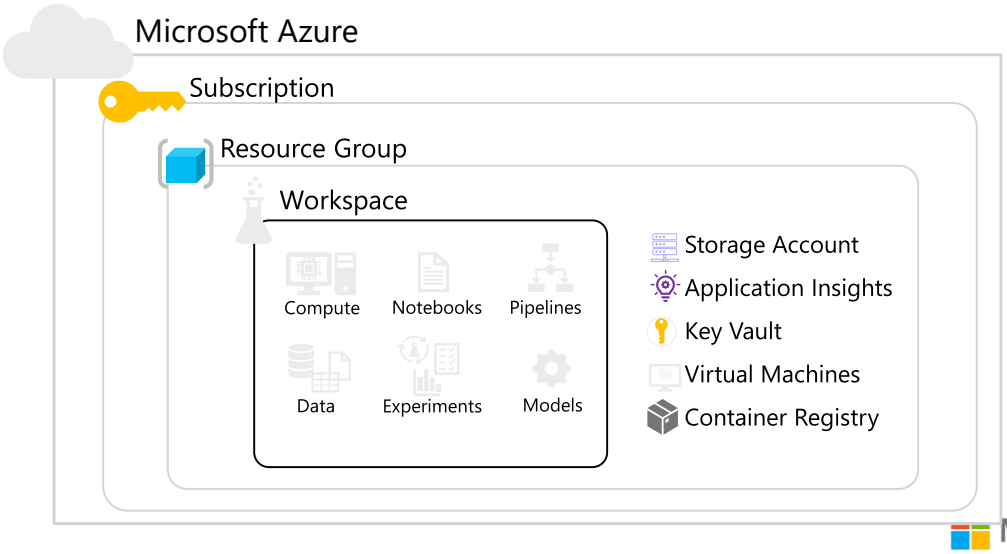
A platform for operating machine learning workloads in the cloud



Microsoft

196

Azure Machine Learning Workspaces



Microsoft

197

Considerations for Creating a Workspace

Region

Check Azure Resource availability

For example, NC-Series Virtual Machines for GPU processing

Edition

Enterprise

- All features

Basic

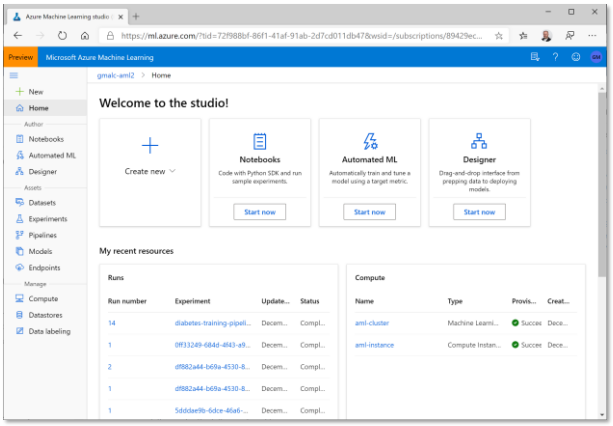
- No Visual Designer
- No Automated ML user interface
- No Data Drift user interface

Microsoft

198

Azure Machine Learning studio

Manage compute and data
Run experiments
View metrics
Manage and deploy models
Manage endpoints
Use graphical modeling tools:
Designer - "no-code" model development
Automated Machine Learning - find the best model for your data



Microsoft

199

The Azure Machine Learning SDK for Python

Code-based configuration for machine learning assets:
Automate repeatable asset creation
Ensure consistency across development, test, and production environments
Incorporate machine learning asset configuration into DevOps

```
pip install azureml-sdk
```

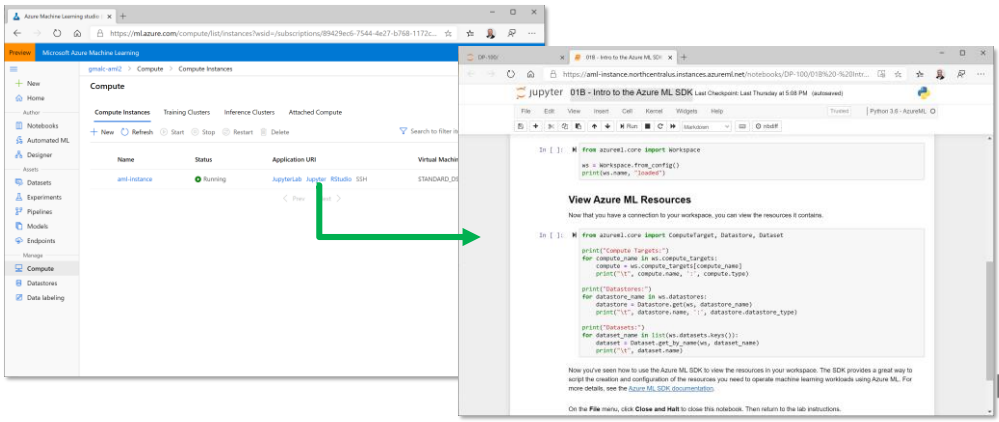
```
from azureml.core import Workspace

ws = Workspace.from_config()
for compute_name in ws.compute_targets:
    compute = ws.compute_targets[compute_name]
    print(compute.name, ":", compute.type)
```

Microsoft

200

Compute Instances

Jupyter Notebook and JupyterLab servers in your workspace
Choose the compute specifications you need



Microsoft

201

Walkthrough :

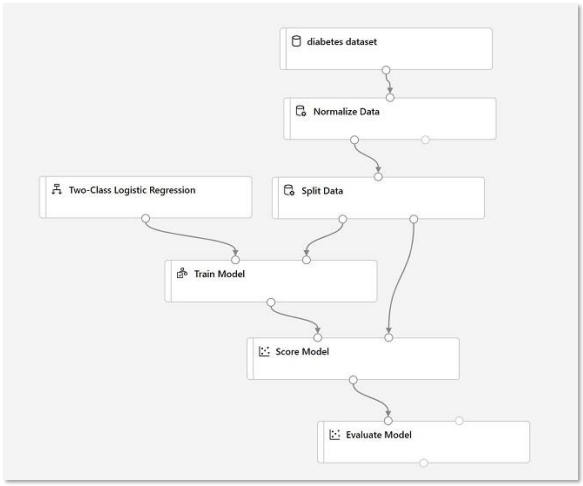
Creating an Azure Machine Learning Workspace

Microsoft

202

What is Azure Machine Learning Designer?

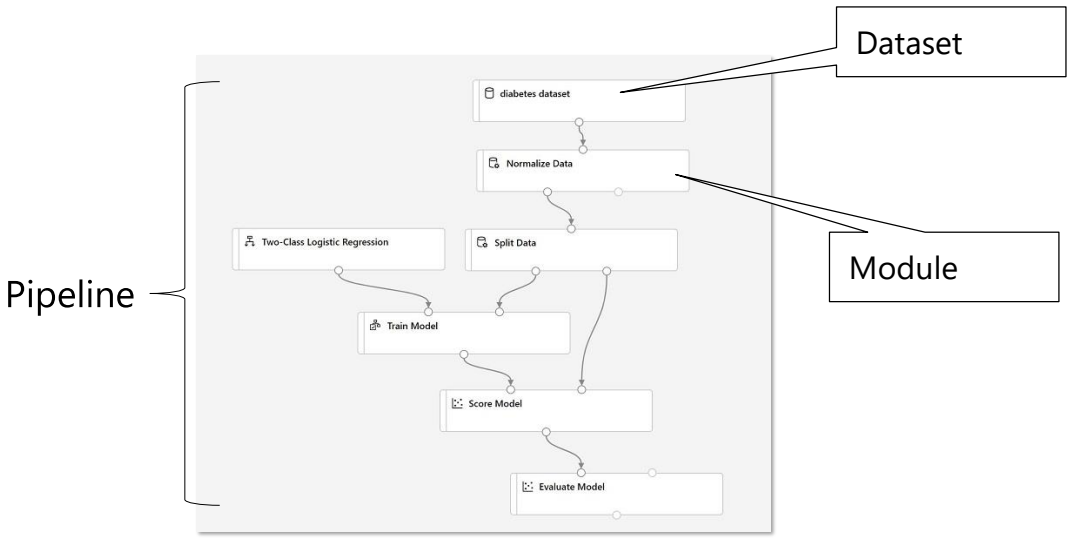
Drag-and-Drop Interface for:
Preparing data and training models
Publishing models as services



Microsoft

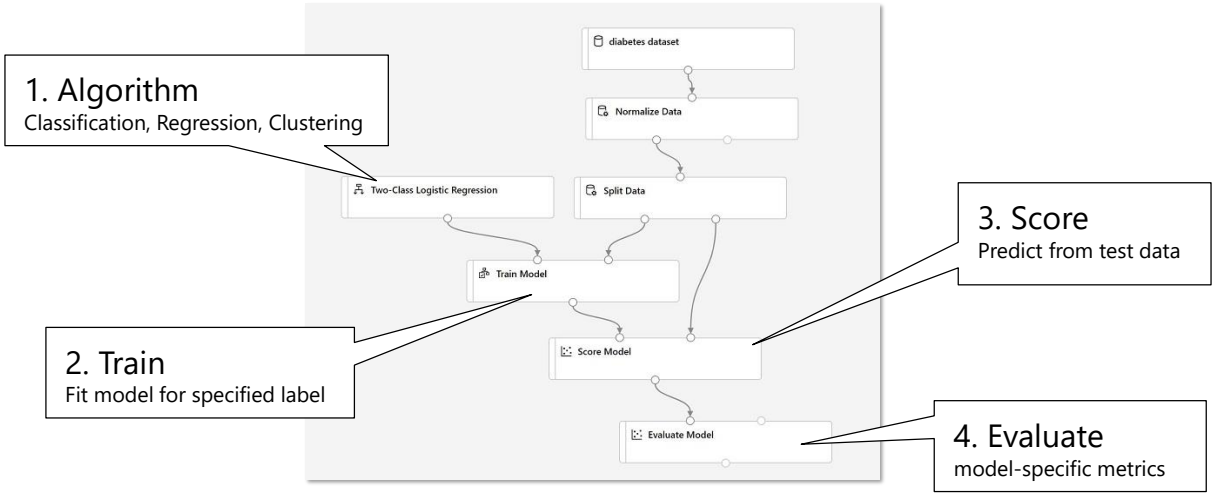
203

Designer Pipelines and Modules



204

Training, Scoring, and Evaluating Models



205

Custom Code Modules

Apply SQL Transformation	Use a SQL statement to process up to three input tables
Execute Python Script	Implement a custom Python function to process up to two dataframes
Create Python Model	Implement a custom Python model in place of a built-in algorithm
Execute R Script	Implement a custom R function to process up to two dataframes

Microsoft

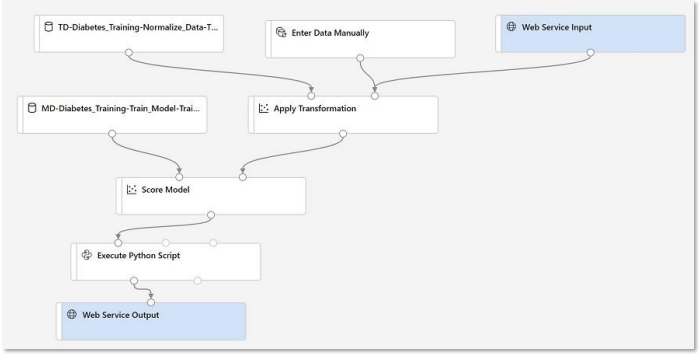
206

Walkthrough :

Creating a Training Pipeline with the Azure ML Designer

Microsoft

What is an Inference Pipeline?



A data flow defining a web service for using the trained model

A **Web Service Input** defines the input data schema
Transformations based on training data are encapsulated in datasets
The trained model is encapsulated in a dataset
A **Web Service Output** defines the output data schema
You may want to modify the pipeline before deploying its as a web service

Microsoft

Publishing a Service Endpoint

Deploy a Real-Time Pipeline:
Requires Azure Kubernetes Services Inference Compute
Submit new data to HTTP endpoint for immediate results

Publish a Batch Pipeline
Requires Azure Machine Learning Training Compute
Initiate pipeline experiment run through HTTP endpoint
Results saved in run output

Microsoft

209

Consuming a Service Endpoint

View endpoints in Azure Machine Learning studio
Use starter code to build client applications

```
data = {"Inputs": {"input0":[{'feature1': "123", 'feature2': "99",},],},
        "GlobalParameters":  {}}
body = str.encode(json.dumps(data))

url = 'http://10.0.0.1:80/api/v1/service/diabetes_predictor/score'
api_key = 'a1234567890x'
headers = {'Content-Type':'application/json',
           'Authorization':('Bearer '+ api_key)}

req = urllib.request.Request(url, body, headers)
response = urllib.request.urlopen(req)
result = response.read()
```

Microsoft

210

Walkthrough :

Deploying a Service with the Azure ML Designer

Microsoft

211

Complete the Course

212