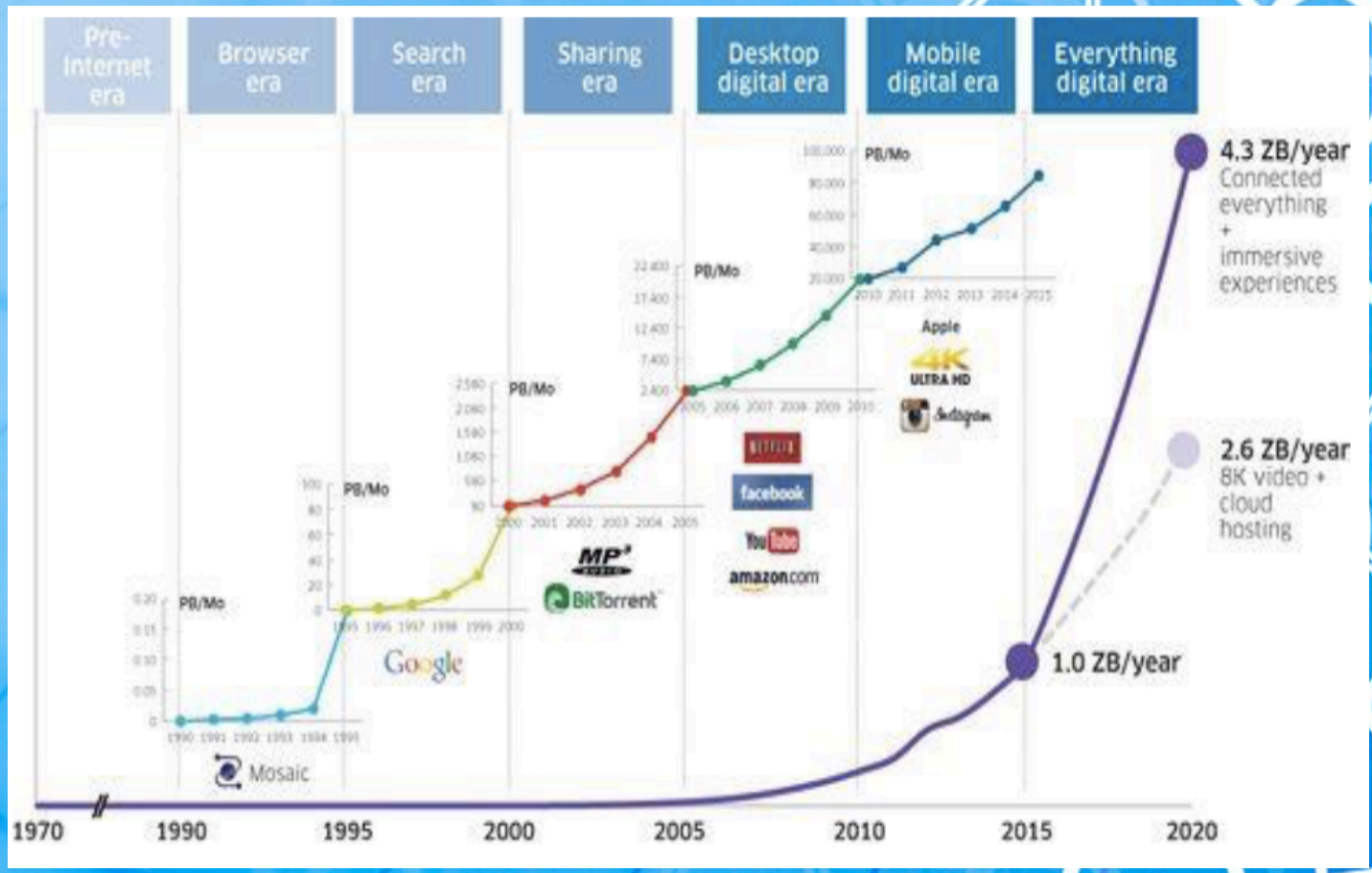

Introduction
to Big Data

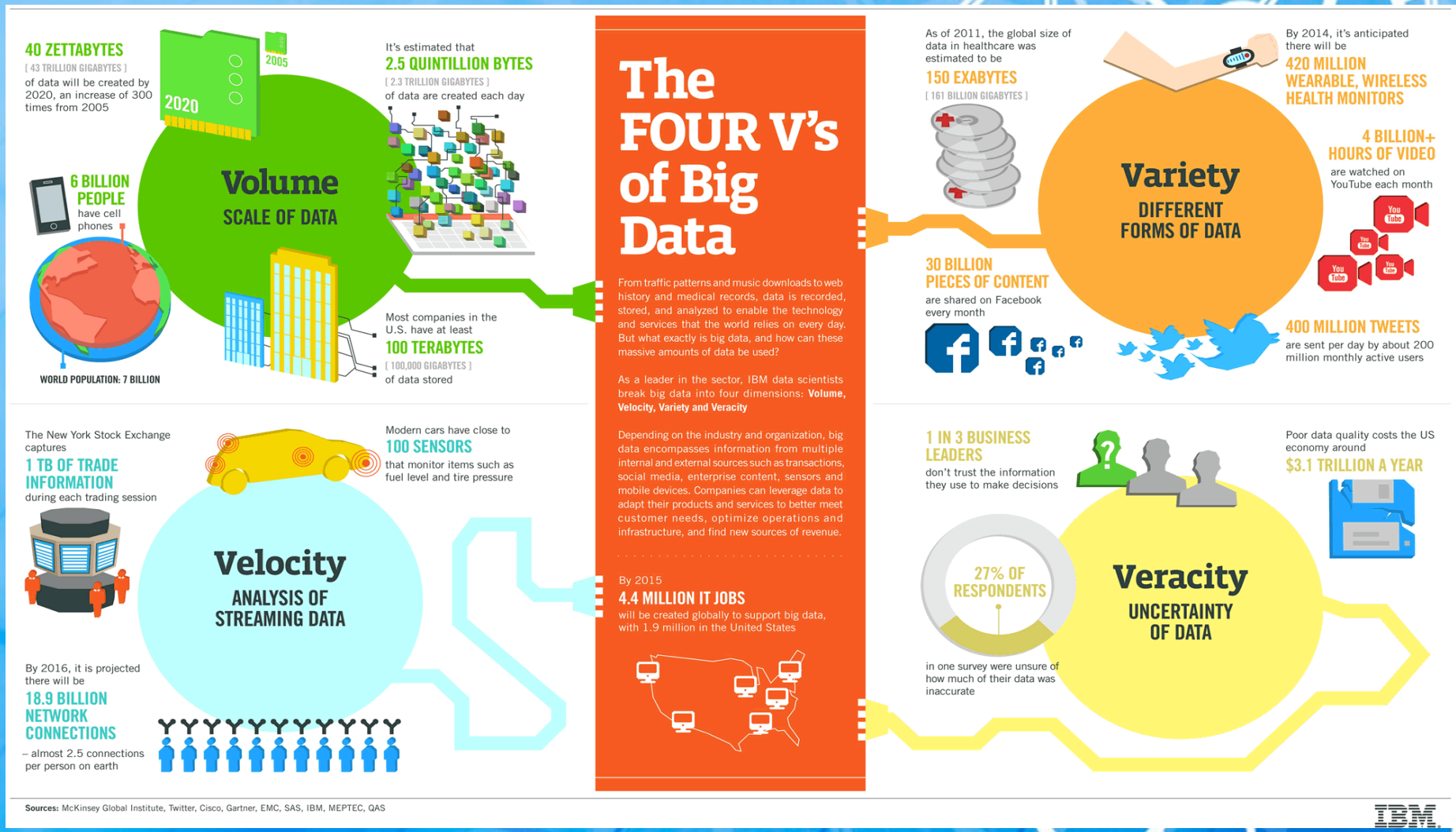
What is Big Data?

Big Data is high- volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

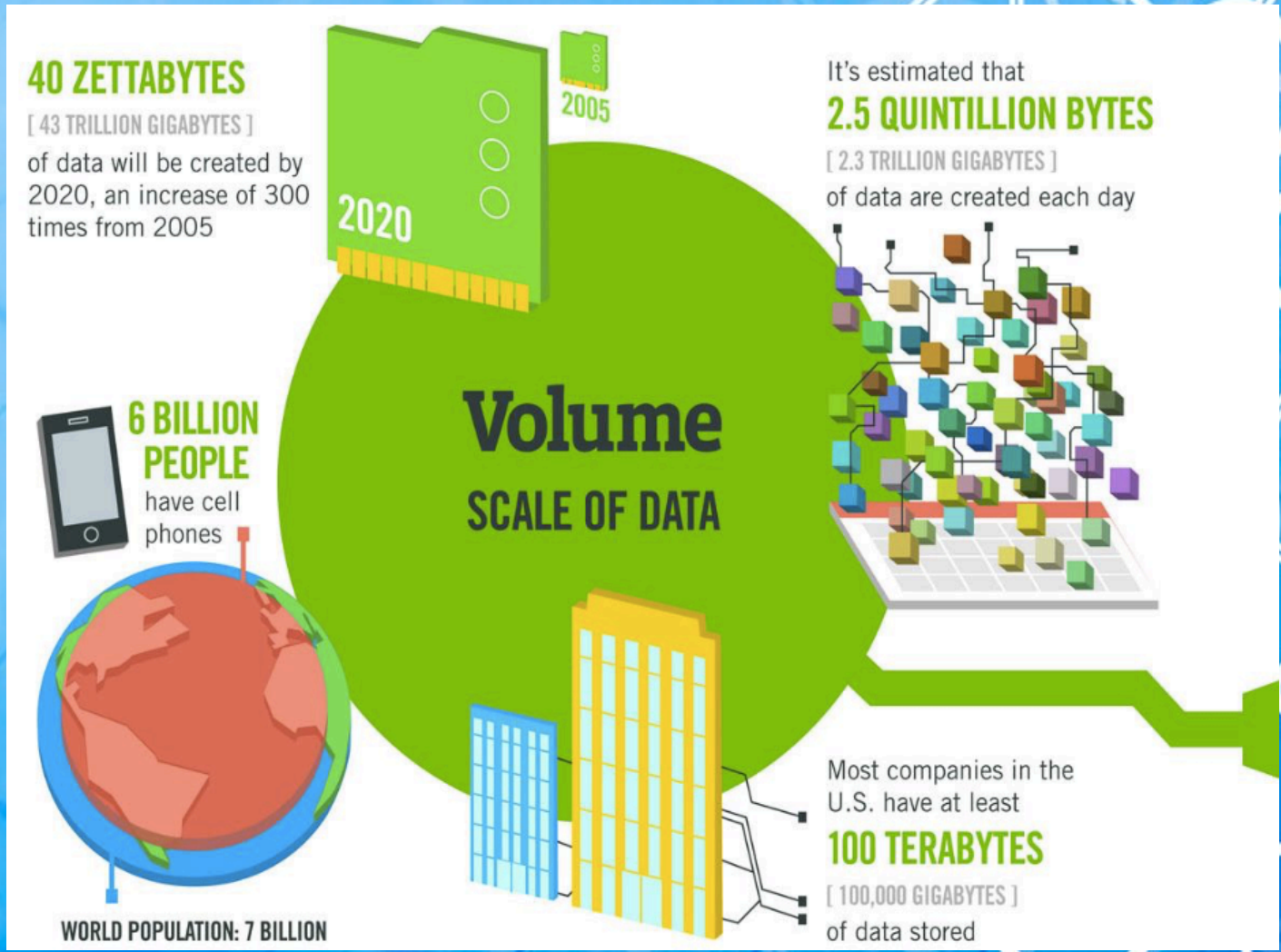
Evolution of Big Data



4v of Big Data



Volume


40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2005
2020
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
Volume
SCALE OF DATA
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Velocity

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session

Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure

Velocity

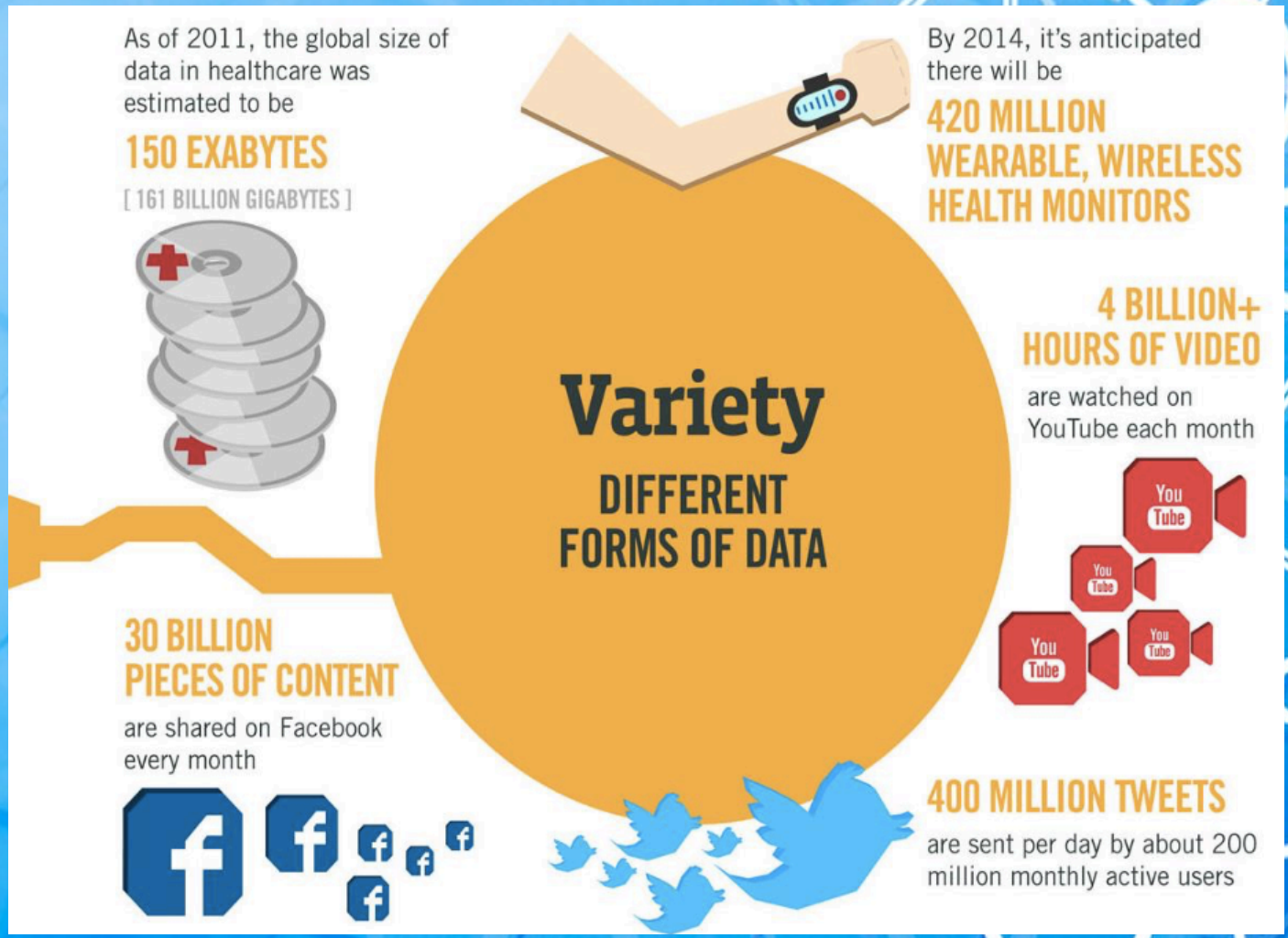
ANALYSIS OF STREAMING DATA

By 2016, it is projected there will be

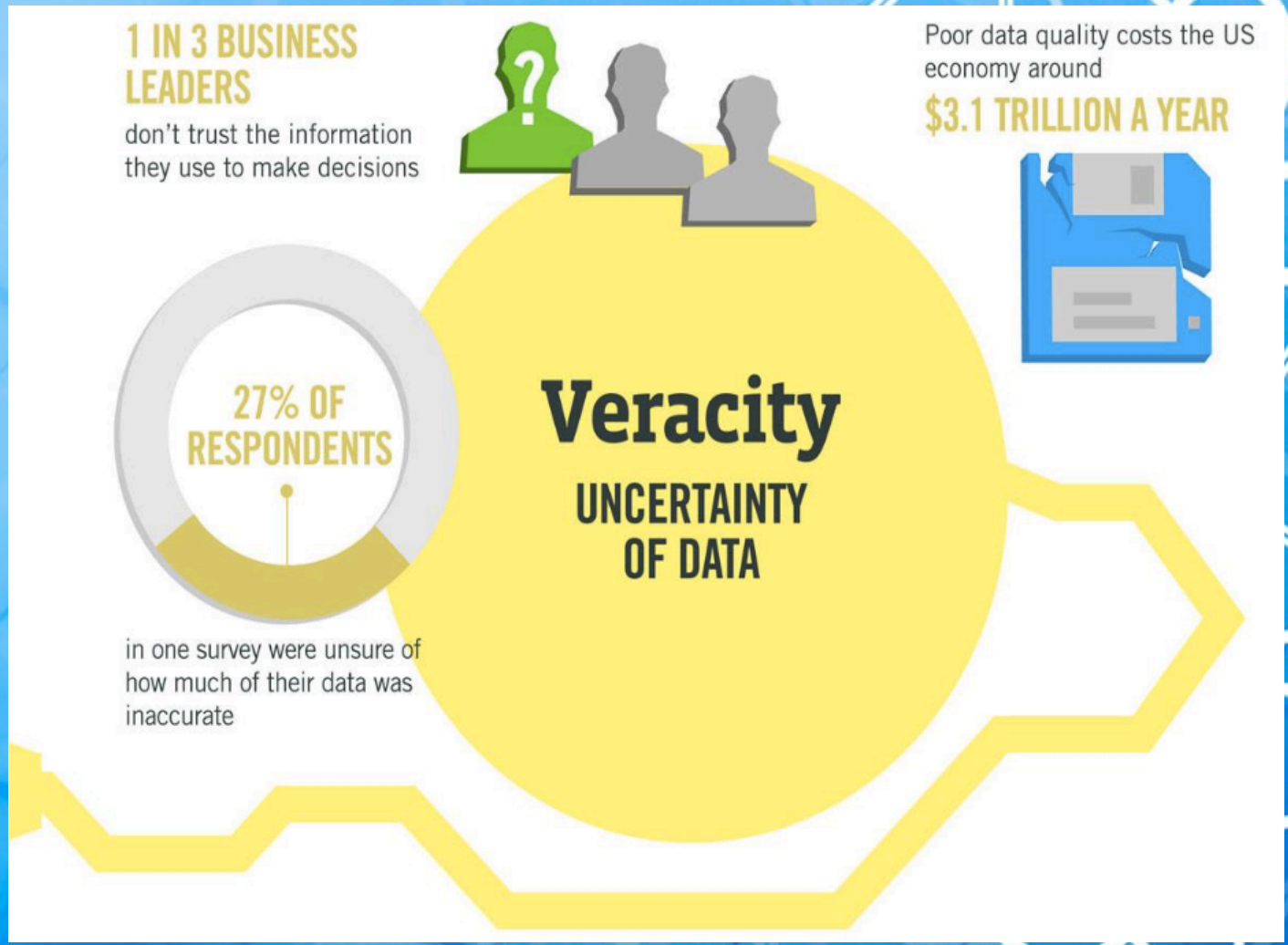
18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth

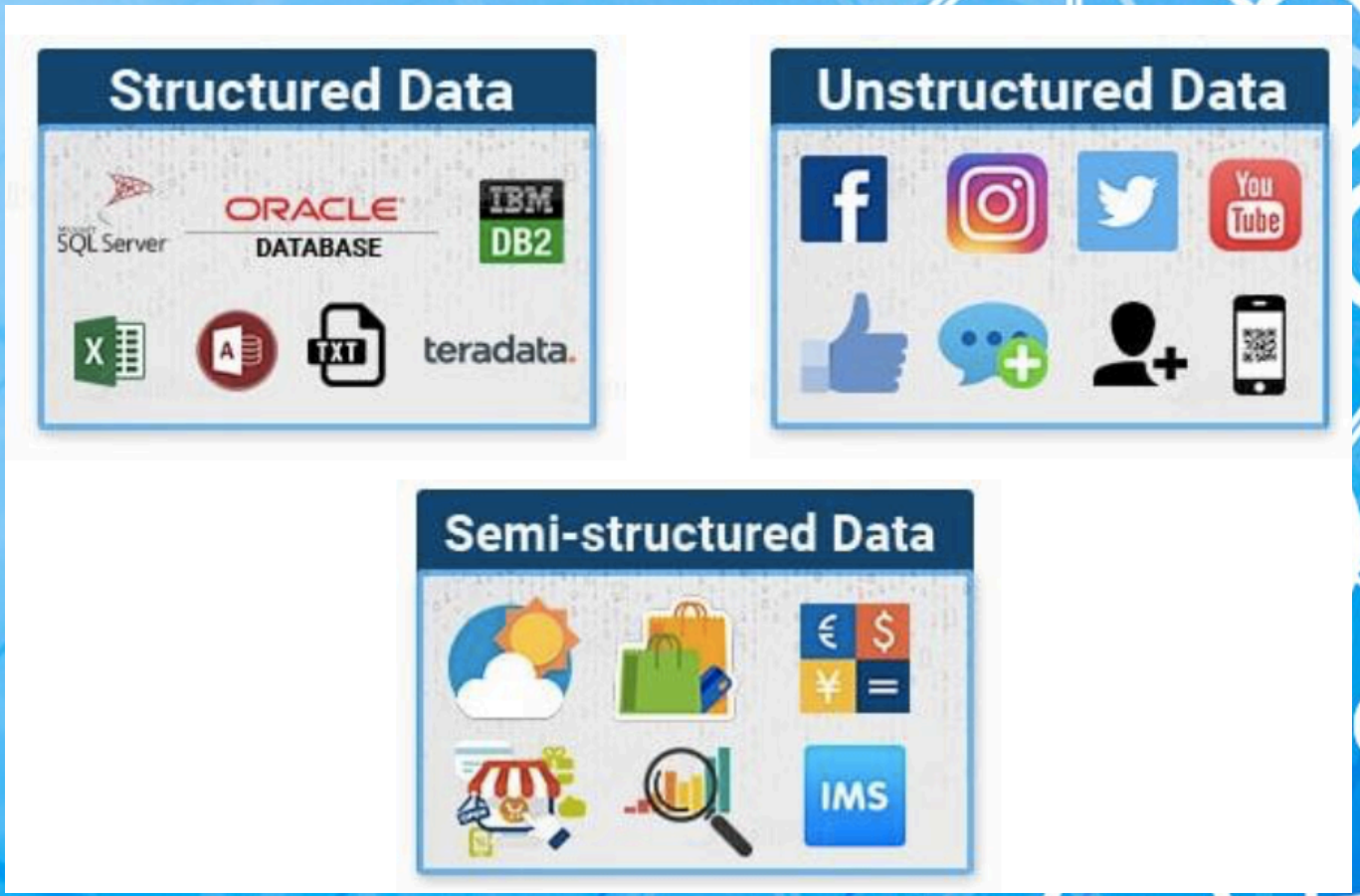
Variety


As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
Variety
DIFFERENT FORMS OF DATA
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users

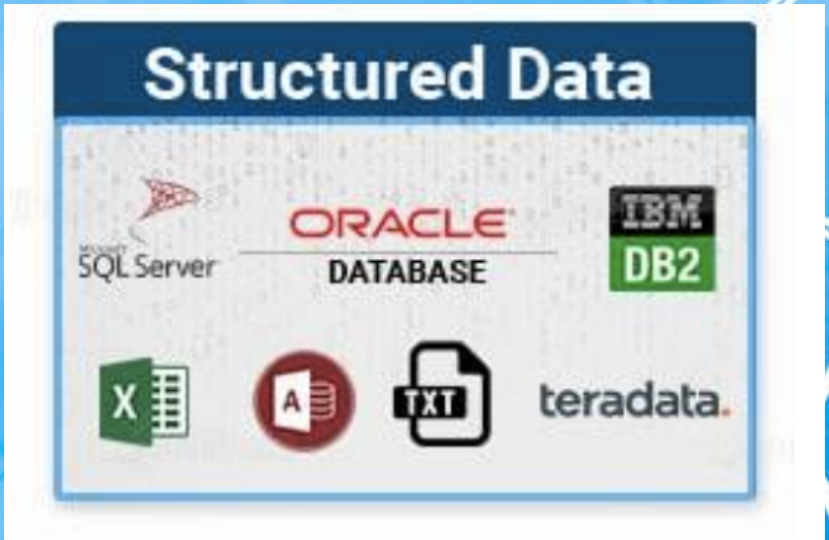
Veracity


1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA

Types of Data



Structured Data



- It can be stored in a tabular column.
- Examples of structured data:

Relational databases, most of the modern computers are able to make sense of structured data.

Unstructured Data

- It can not be fit into tabular databases.
- Examples of unstructured data:

Audio, video, images, emoji's , comments, etc

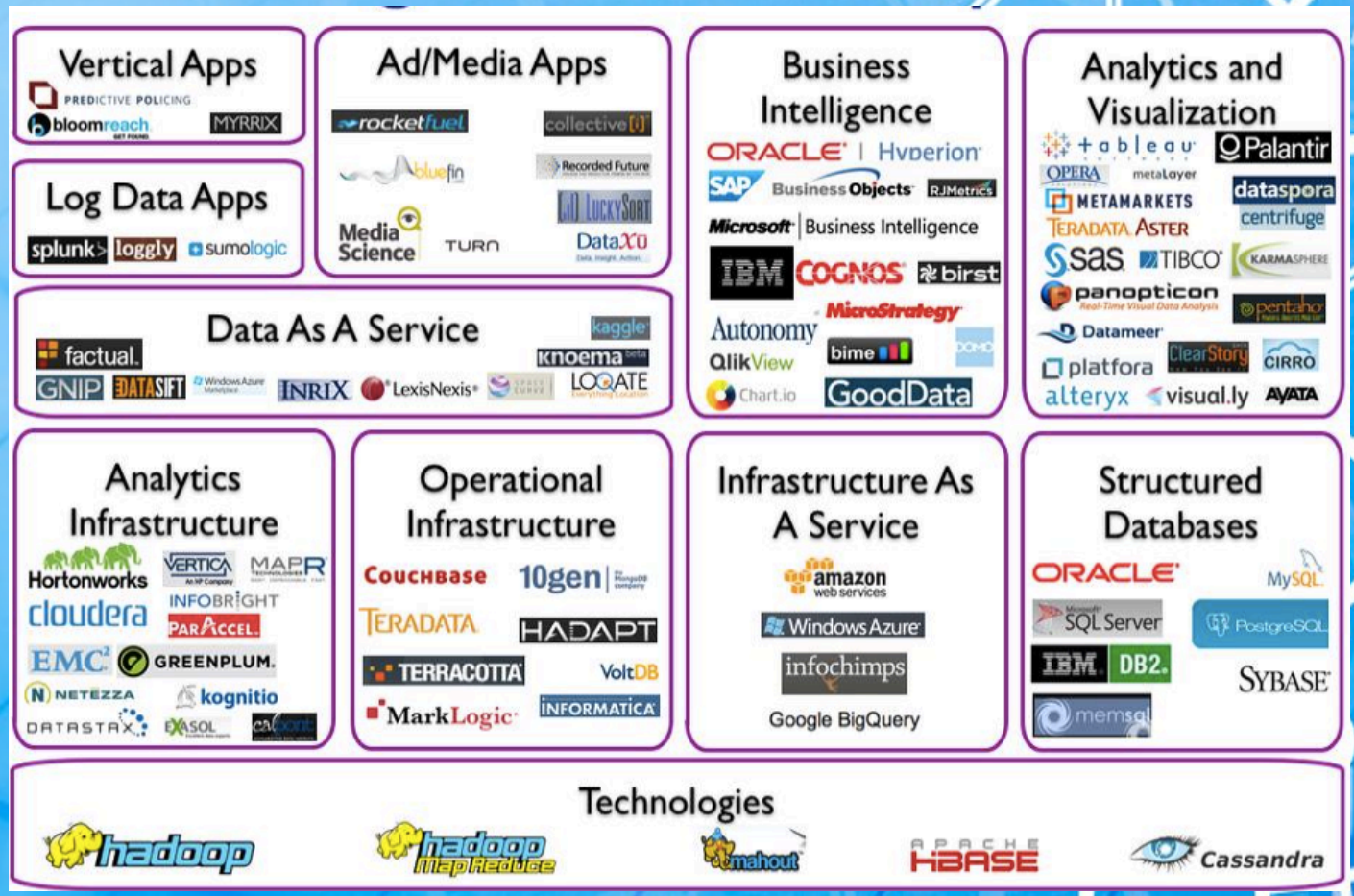
Semi - Structured Data

- It includes both structured and unstructured data

- This type of data sets include a proper structure, but still it might not be possible to sort or process that data due to some constraints.

- Examples of semi-structured data:
XML data, JSON files, and others.

Big Data Landscape



Big Data Technologies

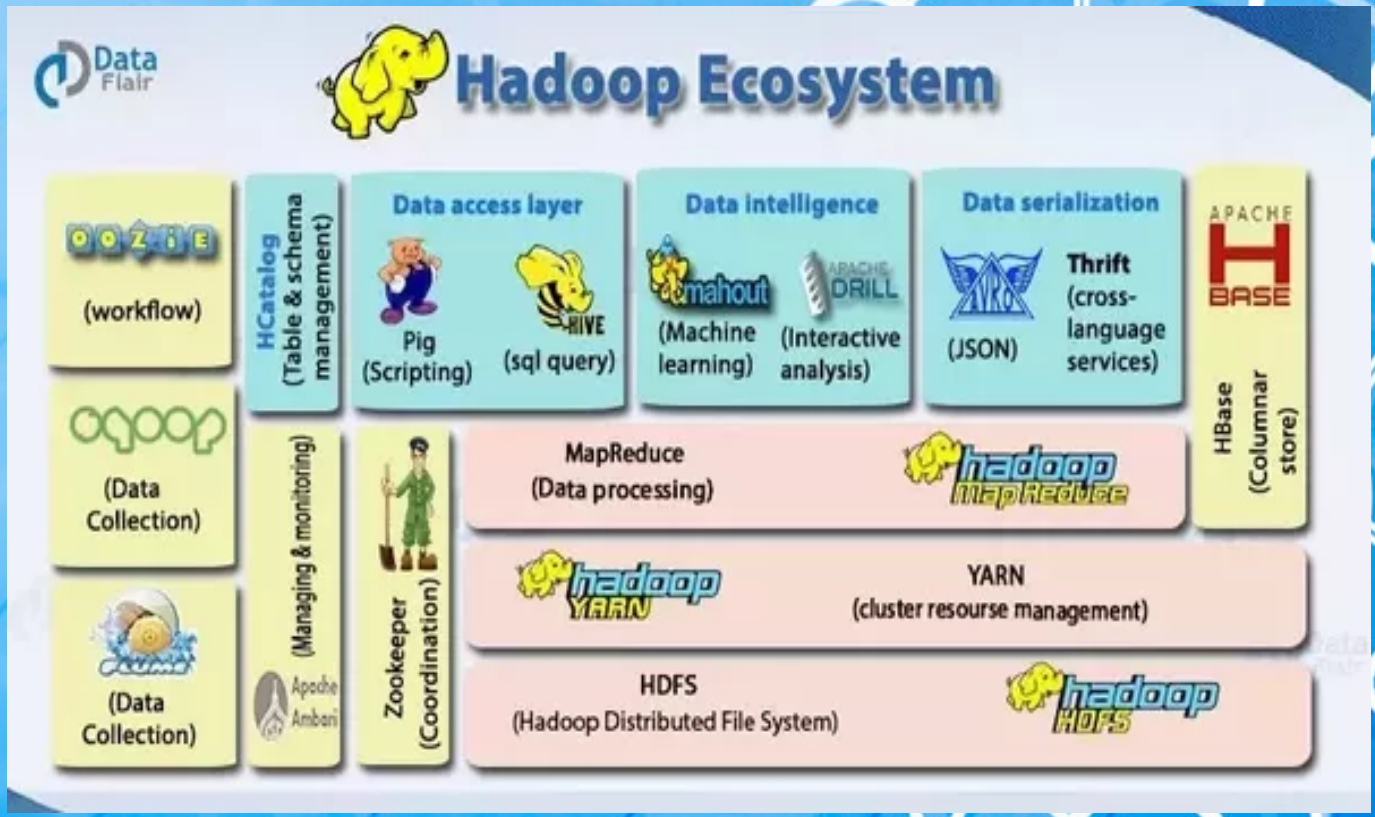
Hadoop

Hadoop is an ecosystem of open source programs and procedures that allows for the distributed processing of large data sets across clusters of computers using simple programming models.

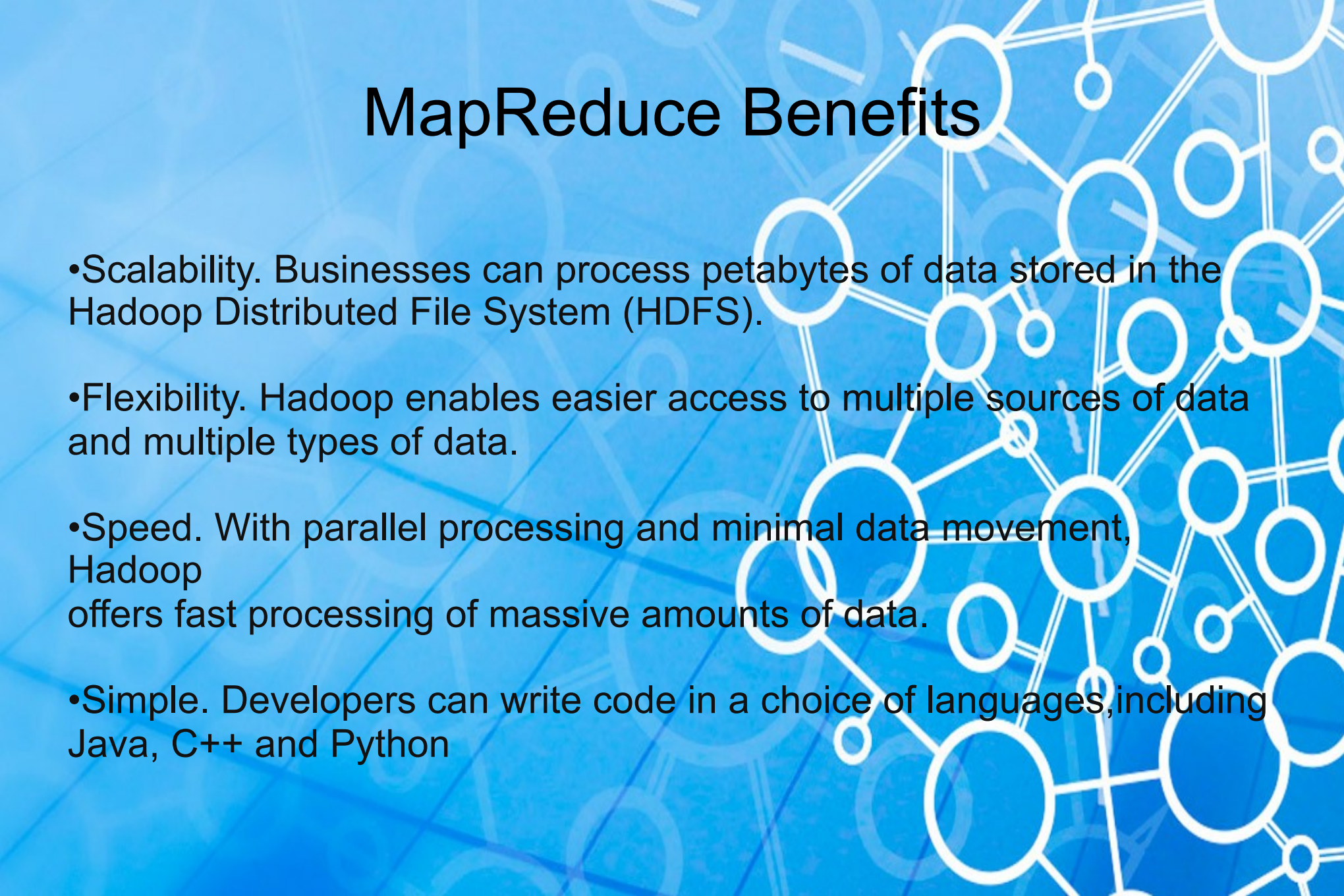
It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.

Hadoop used by

Hadoop Ecosystem



MapReduce Benefits



•Scalability. Businesses can process petabytes of data stored in the Hadoop Distributed File System (HDFS).

•Flexibility. Hadoop enables easier access to multiple sources of data and multiple types of data.

•Speed. With parallel processing and minimal data movement, Hadoop
offers fast processing of massive amounts of data.

•Simple. Developers can write code in a choice of languages,including Java, C++ and Python

The Cloud

Requirement

The XYZ company is implementing a solution for getting more knowledge of his API use and network performance. For this propose, it has requested to us a nonintrusive solution for sniffing the traffic between clients and server. Around 300.000.000 of requests to API are done every day.

The company wants to achieve this objective:

- Dashboards to monitor in real time the performance of the network. We get information of the HTTP package headers. The data we capture here is Latency, HTTP error code, size of the binary, IP source, and IP destiny. Here we need to work in real time and data cannot be aggregated. This database will be synchronized with an alarming system, so can be read and write action simultaneously. We have to retain the information for 1 week without aggregation, then, we will at a level of 5 minutes.

- Dashboards to analyze the API's request and responses payload that contains values information. Datumize's software transforms the XML to an object that is inserted into the database. At this point, we are grouping the data on periods of 5 minutes and we had a column with the number of occurrences. When the day passes, this information is grouped in blocks of hours. When a week is done, the previous week is grouped by days. This information could have a delay of 2 hours.

Requirement

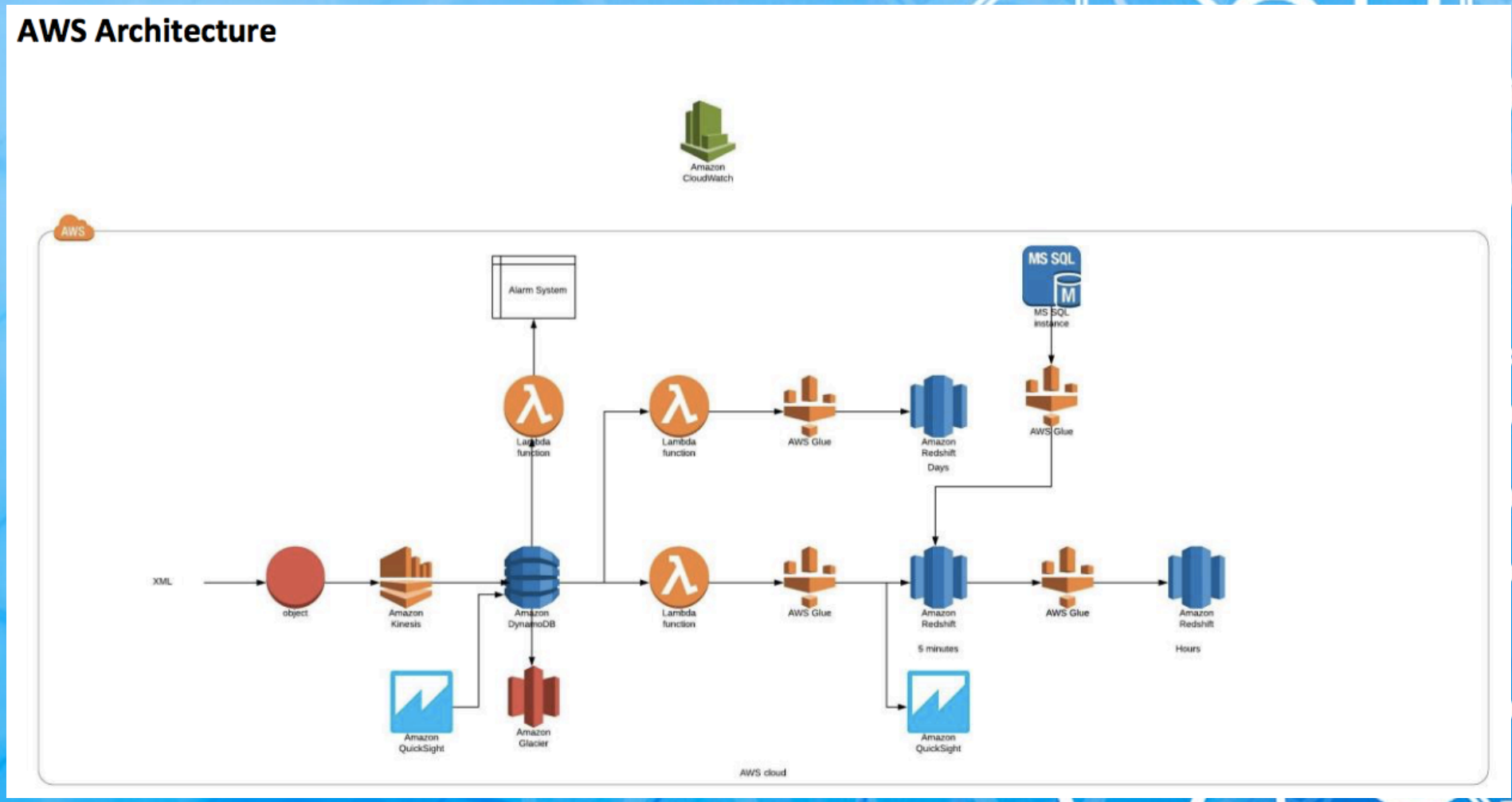
- The raw information has to be retained on a file system for a week. No grouping at this level.

There is something else to keep on the mind. On the database, one of the fields is a list of Id's of elements. This ID's are separated by '#', for example, 34983489 # 939343894 # 298289289 . One of the statistics we want to display on our reporting is filtering by Id, so we would need to disaggregate the information somehow to optimize this queries.

That's not all. To get an alphanumeric to the ID's described previously we synchronize with a SQL database we had to connect via SSH. The amount of data to synchronize is of 20.000.000 of rows, that can be updated, but we do the assumption that we need to import the daily increment. Once a month we run a drop and insert all data process.

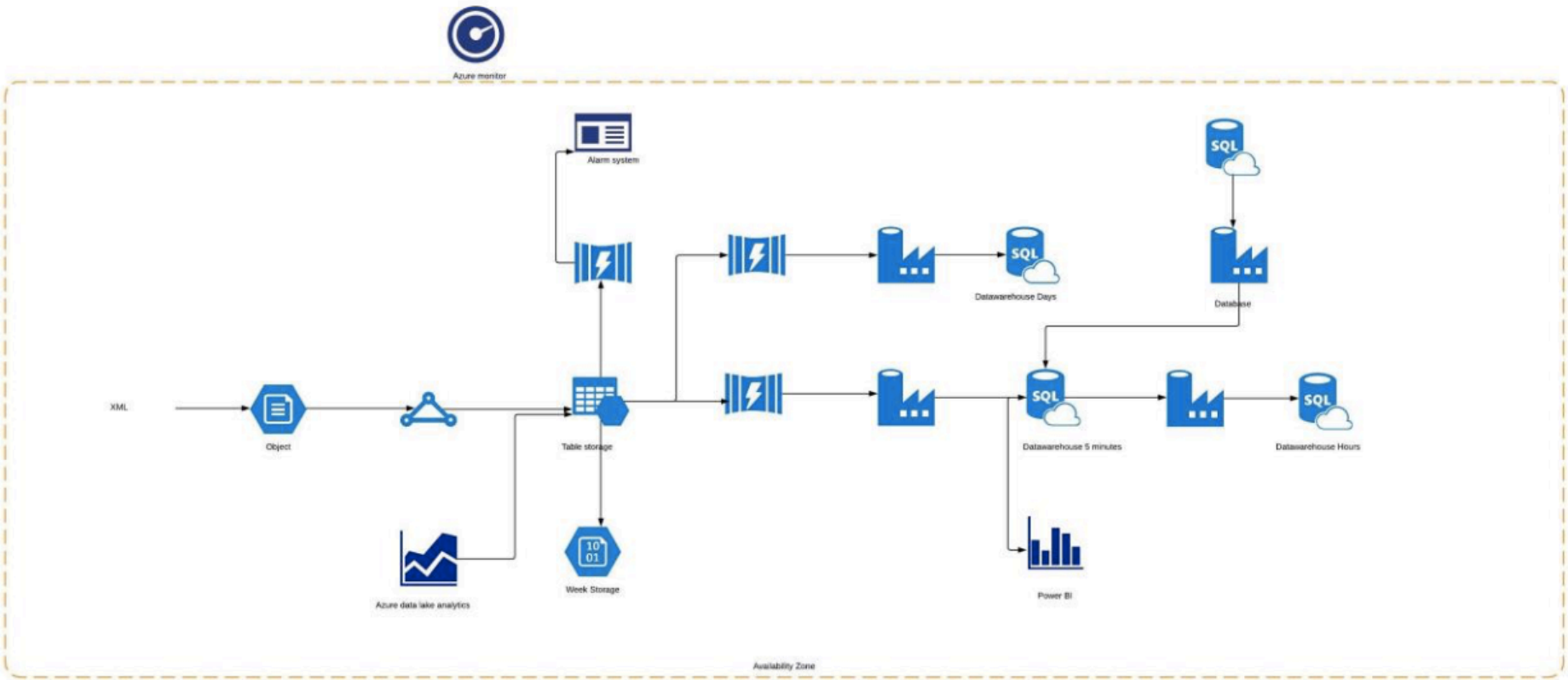
We request you to define two architecture solution, one of them using Amazon infrastructure and another at your choice. Also, explain all the process involved and the technology decided to solve it. Add a description of the tables inside the database. You don't have to provide a solution for the dashboards. Also, add the disadvantages of your solutions.

AWS

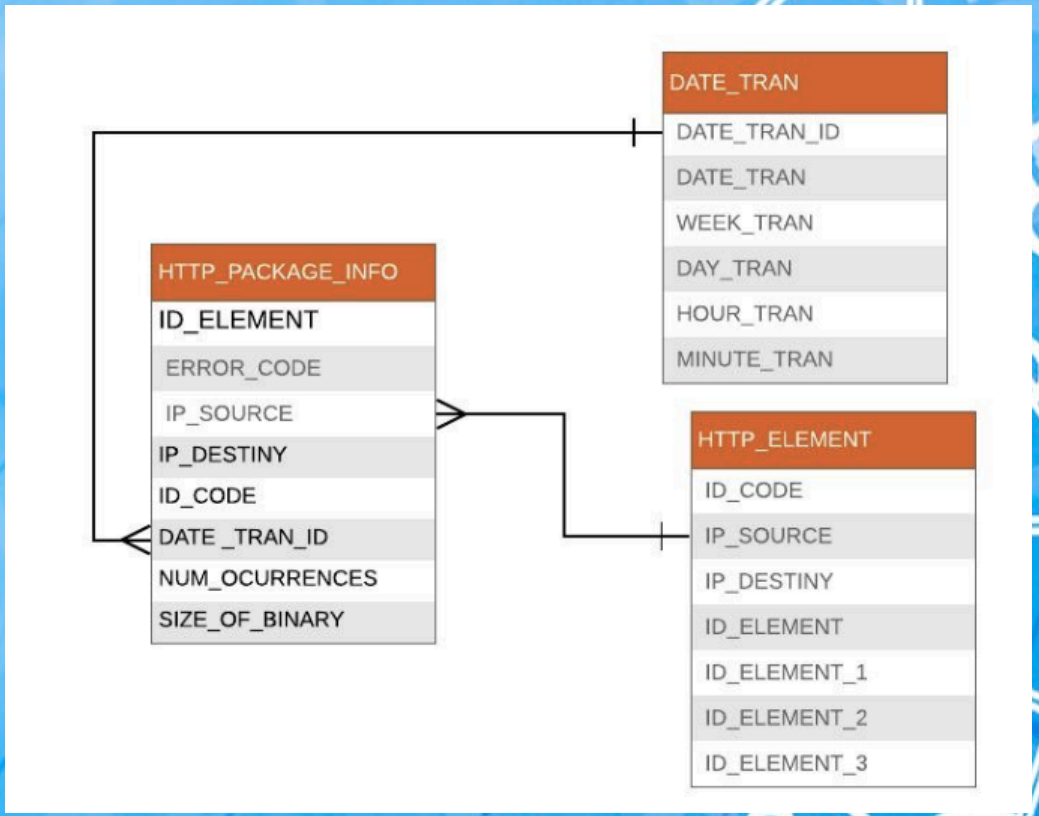

AWS Architecture
Amazon CloudWatch
AWS
Alarm System
MS SQL
MS SQL instance
Lambda function
Lambda function
AWS Glue
Amazon Redshift
Days
AWS Glue
XML
object
Amazon Kinesis
Amazon DynamoDB
Lambda function
AWS Glue
Amazon Redshift
5 minutes
AWS Glue
Amazon Redshift
Hours
Amazon QuickSight
Amazon Glacier
Amazon QuickSight
AWS cloud

Azure

Azure Architecture



Datawarehouse



Tables

Granularity: The lowest level is in minutes

HTTP_PACKAGE_INFO: Fact Table that refer to the principal information about http package

ID	COLUMN	DESCRIPTION
1	ID_ELEMENT	ELEMENT COMPUEST OF THE HEADER
2	ERROR_CODE	CODE OF ERROR
3	IP_SOURCE	IP ORIGIN
4	IP_DESTINY	IP DESTINY
5	ID_CODE	ID OF THE TABLE ELEMENT
6	DATE_TRAN_ID	ID OF THE DATE
7	NUM_OCURRENCES	NUMBER OF OCURRENCES
8	SIZE_OF_BINARY	SIZE OF BINARY

HTTP_ELEMENT: Dimension that store information about the element of the package in disaggregated level

Tables

ID	COLUMN	DESCRIPTION
1	ID_CODE	ID OF THE TABLE ELEMENT
2	IP_SOURCE	IP ORIGIN
3	IP_DESTINY	IP DESTINY
4	ID_ELEMENT	ELEMENT COMPUEST OF THE HEADER
5	ID_ELEMENT_1	ID ELEMENT 1
6	ID_ELEMENT_2	ID ELEMENT 2
7	ID_ELEMENT_3	ID ELEMENT 3

DATE_TRAN: Dimension that refer to the date of transaction

ID	COLUMN	DESCRIPTION
1	DATE_TRAN_ID	CODE OF THE TABLE DATE_TRAN
2	DATE_TRAN	DATE OF THE TRANSACTION
3	WEEK_TRAN	WEEK OF THE TRANSACTION
4	DAY_TRAN	DAY OF TRNSACTION
5	HOUR_TRAN	HOUR OF TRANSACTION
6	MINUTE_TRAN	MINUTE OF TRANSACTION

Jobs

- Data Engineer
- Data Scientist
- Machine Learning Engineer
- Business Analyst
- Data Architect

Thanks