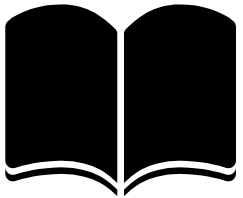
BIG DATA AND NoSQL

LEARNING OBJECTIVES



WHAT BIG DATA IS AND WHY IT IS IMPORTANT IN MODERN BUSINESS

THE PRIMARY CHARACTERISTICS OF BIG DATA AND HOW THESES GO BEYOND THE TRADITIONAL “3 V’s”

THE FOUR MAJOR APPROACHES OF THE NoSQL MODEL AND HOW THEY DIFFER FROM THE RELATIONAL MODEL

ABOUT DATA ANALYTICS,INCLUDING DATA MINING AND PREDICTIVE ANALYTICS

DISCUSSING AND REVIEWING POWERBI

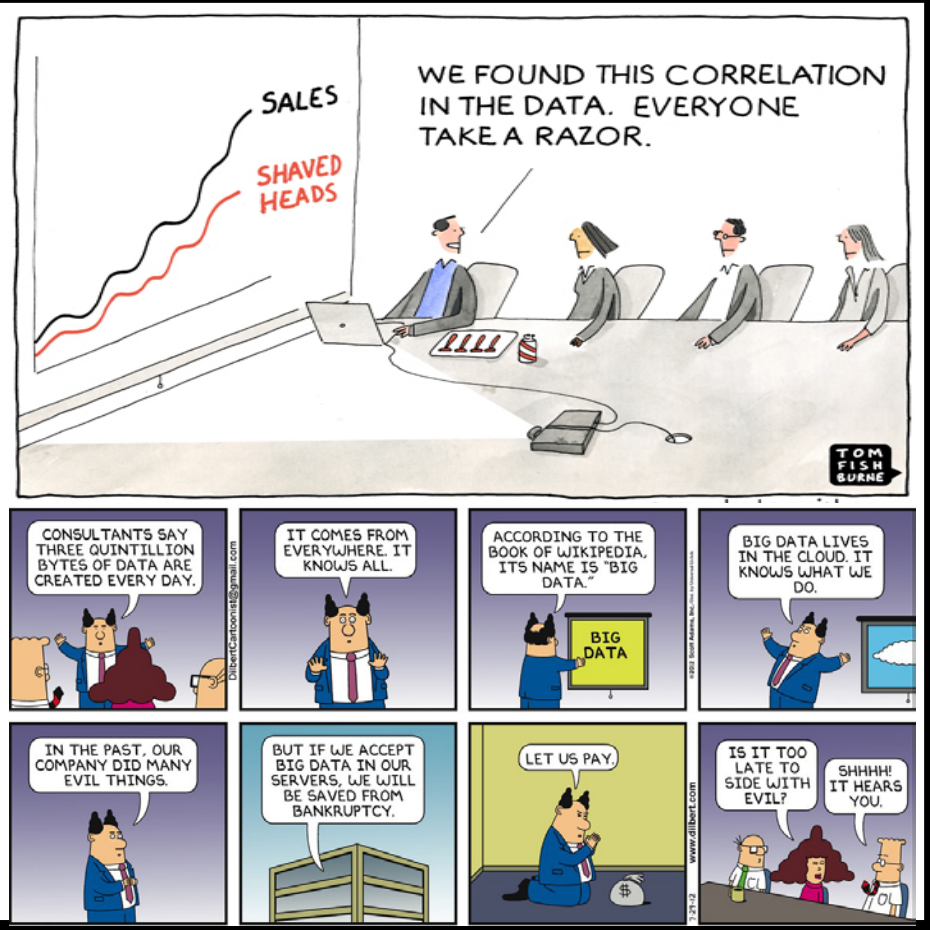
REVIEWING NoSQL DOCUMENT DATABASE – MONGODB & COUCHDB

SAMPLE APPLICATION USING MONGODB

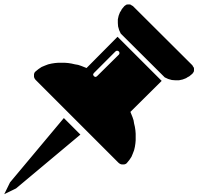
BIG DATA

What is Big Data?

extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions.



BIG DATA



Volume: Quantity of data to be stored

Scaling up is keeping the same number of systems but migrating each one to a larger system

- Ex. Change from a server with 16 CPU cores and a 1 TB storage system to a server with 64 CPU cores and a 100 TB storage system

Scaling out means when the workload exceeds server capacity, it is spread out across a number of servers

- Ex. Also referred to as clustering – creating a cluster of low-cost servers to share the workload.

Velocity: Speed at which data is entered into system and must be processed

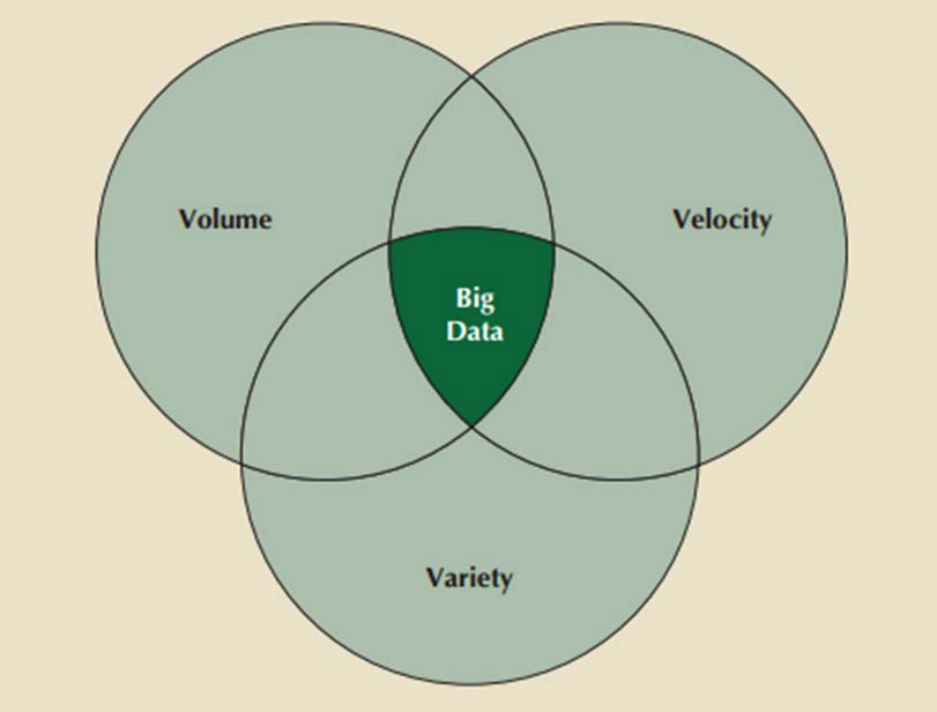
- **Stream processing** focuses on input processing and requires analysis of data stream as it enters the system
- **Feedback loop processing** refers to the analysis of data to produce actionable results

Variety:

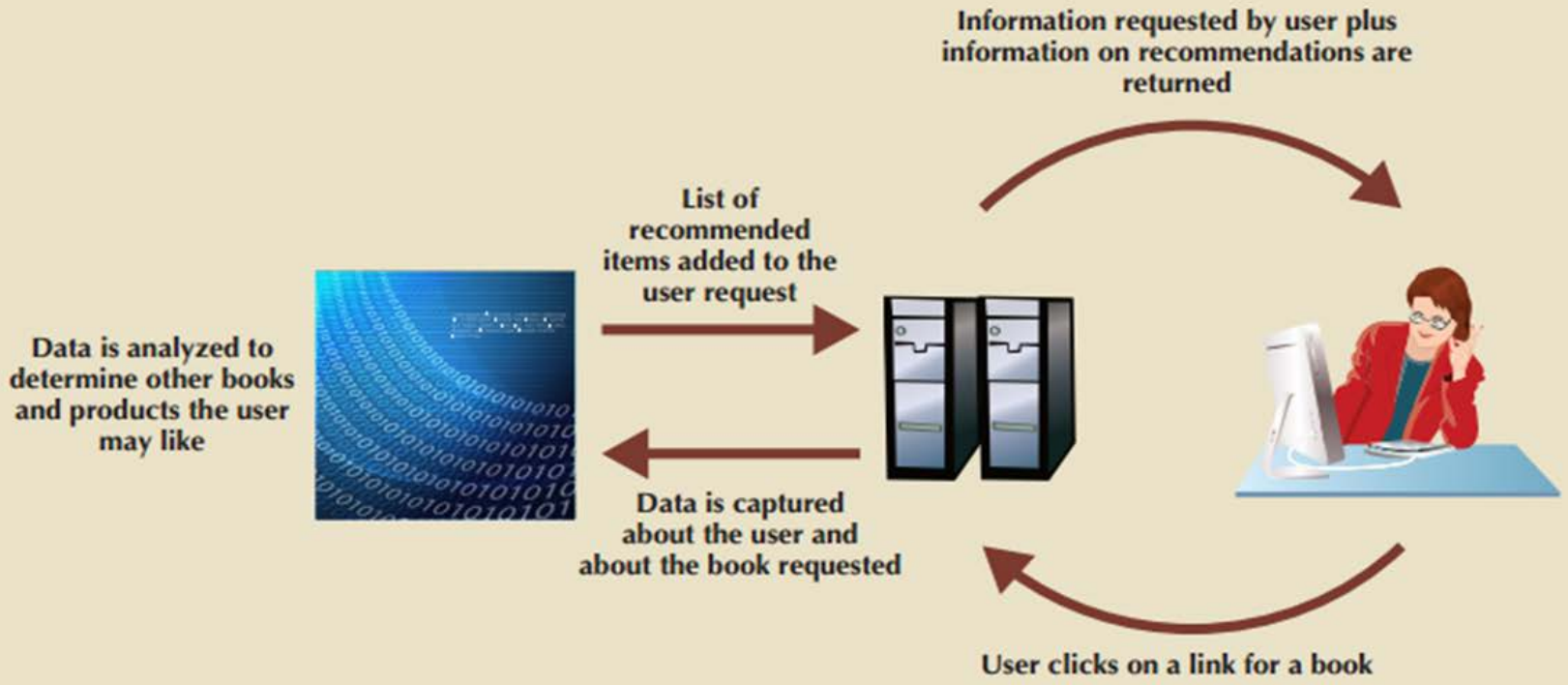
Variations in the structure of data to be stored

- **Structured data** fits into a predefined data model
- **Unstructured data** does not fit into a predefined model

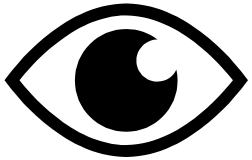
Current View of Big Data



Feedback Loop Processing



BIG DATA



VERACITY
Trustworthiness of data

VARIABILITY
Variability: Changes in meaning of data based on context
- **Sentimental analysis** attempts to determine attitude

VERSATILE
- Characteristics important in working with data in relational models are universal and also apply to Big Data
- Relational databases not necessarily best for storing and managing all organizational data

VALUE
Degree data can be analyzed for meaningful insight

VISUALIZATION
Ability to graphically present data to make it understandable to users

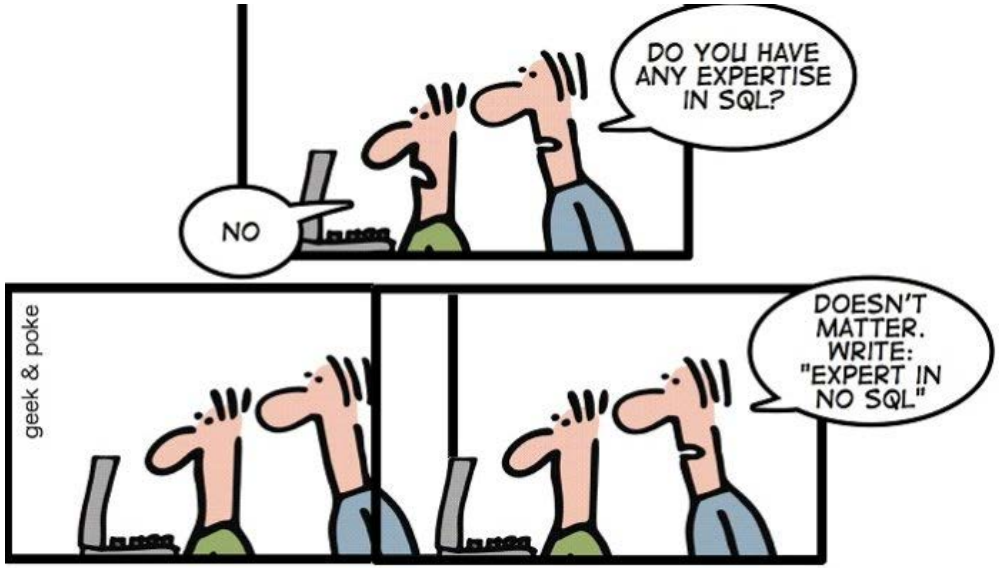
POLYGLOT PERSISTENCE

Coexistence of a variety of data storage and management technologies within an organization's infrastructure

- Most companies are moving to this structure
- https://www.youtube.com/watch?v=1k64HZq28Wc

NoSQL

Name given to non-relational database technologies developed to address Big Data challenges

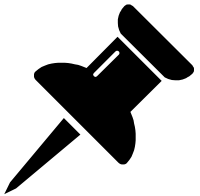


NoSQL DATABASES

BELOW IS THE URL TO A SOURCE THAT RANKS DATABASES
https://db-engines.com/en/ranking

NoSQL Category	Example Databases
Key-value database	Dynamo Riak Redis Voldemort
Document databases	MongoDB CouchDB (Cloudant – cloud hosted version) OrientDB RavenDB
Column-oriented databases	Hbase Cassandra Hypertable
Graph Databases	Neo4J ArangoDB GraphBase

NoSQL



Name given to non-relational database technologies developed to address Big Data challenges

- **Key-value (KV) databases** store data as a collection of key-value pairs organized as buckets which are the equivalent of tables
- **Document databases** store data in key-value pairs in which the value components are tag-encoded documents grouped into logical groups called **collections**

Key-Value Database Storage

Bucket = Customer

Key	Value
10010	"LName Ramas FName Alfred Initial A Areacode 615 Phone 844-2573 Balance 0"
10011	"LName Dunne FName Leona Initial K Areacode 713 Phone 894-1238 Balance 0"
10014	"LName Orlando FName Myron Areacode 615 Phone 222-1672 Balance 0"

Document Database Tagged Format

Collection = Customer

Key	Document
10010	{LName: “Ramas”, FName: “Alfred”, Initial: “A”, Areacode: “615”, Phone: “844-2573”, Balance: “0”}
10011	{LName: “Dunne”, FName: “Leona”, Initial: “K”, Areacode: “713”, Phone: “894-1238”, Balance: “0”}
10014	{LName: “Orlando”, FName: “Myron”, Areacode: “615”, Phone: “222-1672”, Balance: “0”}

NoSQL

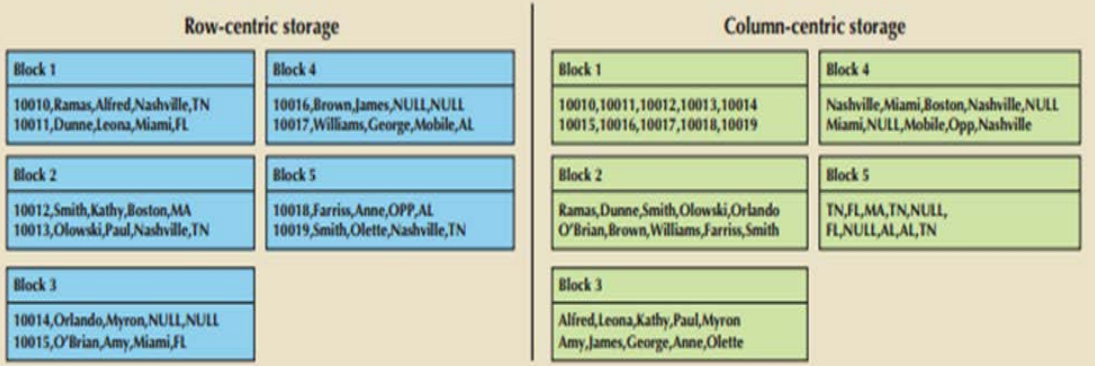
- **Column-oriented databases** refers to two technologies:
  - **Column-centric storage:** Data stored in blocks which hold data from a single column across many rows
  - **Row-centric storage:** Data stored in block which hold data from all columns of a given set of rows

- **Graph databases** store data on relationship-rich data as a collection of **nodes** and **edges**
  - **Properties** are the attributes of a node or edge of interest to a user
  - **Traversal** is a query in a graph database

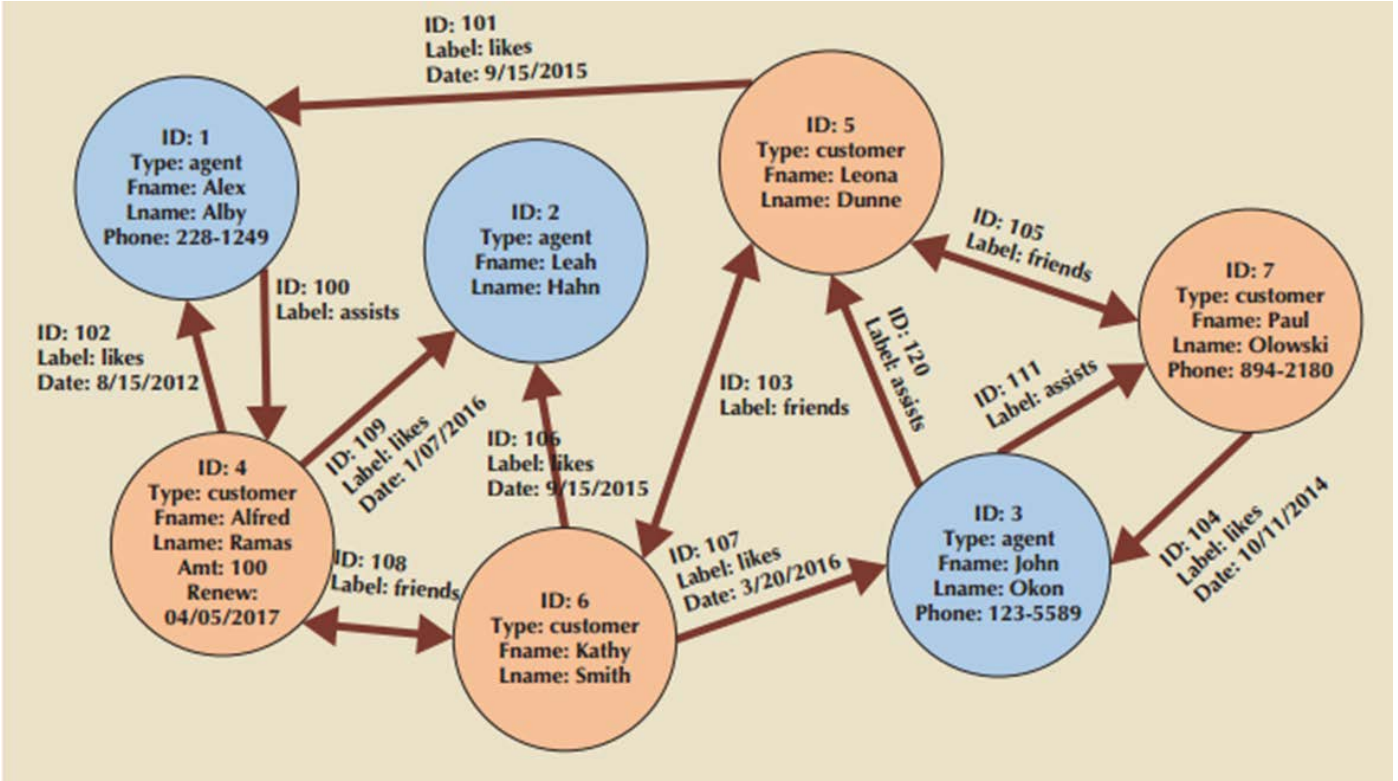
Comparison of Row-Centric and Column-Centric Storage

CUSTOMER relational table

Cus_Code	Cus_LName	Cus_FName	Cus_City	Cus_State
10010	Ramas	Alfred	Nashville	TN
10011	Dunne	Leona	Miami	FL
10012	Smith	Kathy	Boston	MA
10013	Olowski	Paul	Nashville	TN
10014	Orlando	Myron		
10015	O'Brian	Amy	Miami	FL
10016	Brown	James		
10017	Williams	George	Mobile	AL
10018	Farriss	Anne	Opp	AL
10019	Smith	Olette	Nashville	TN



Graph Database Representation



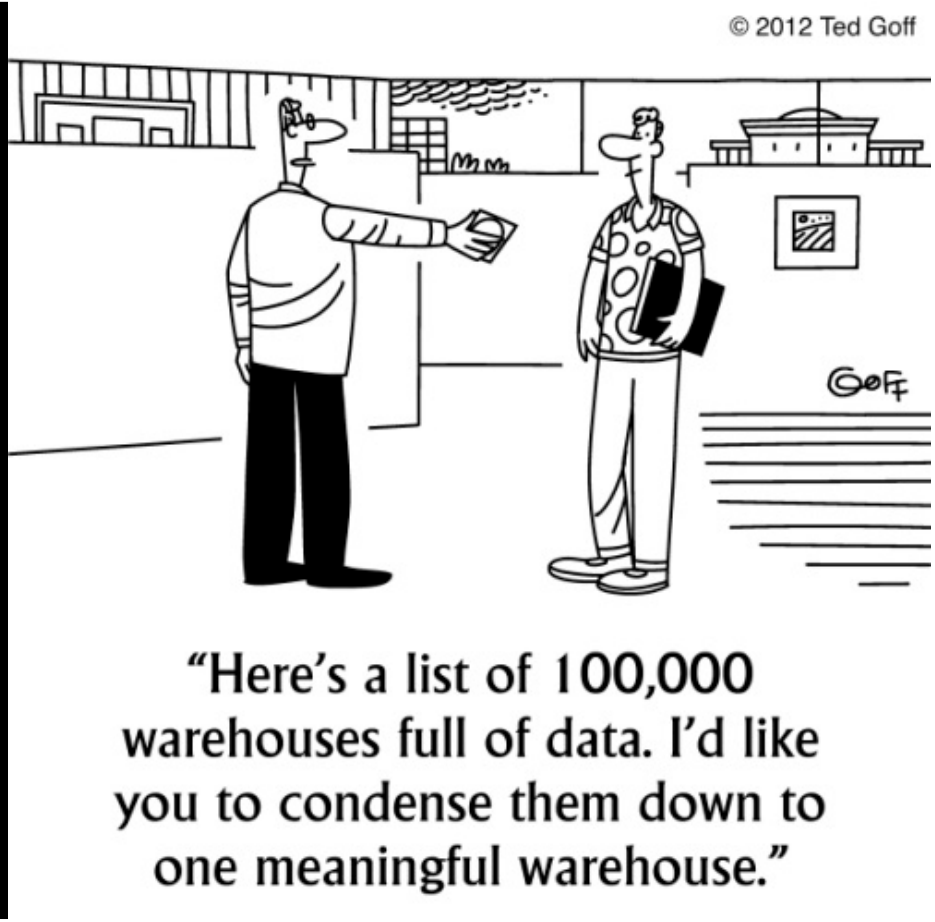
New SQL Databases

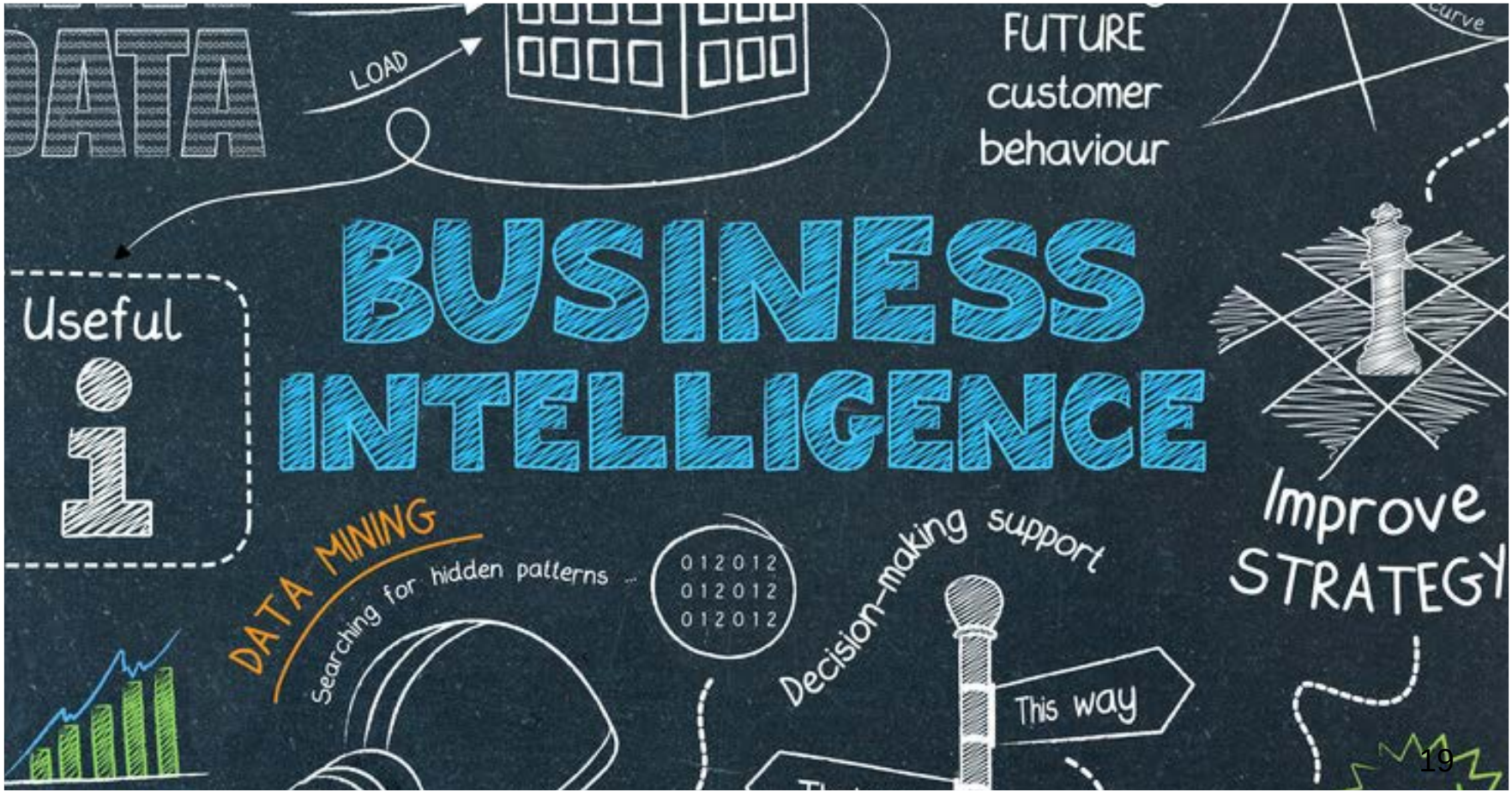
Database model that attempts to provide ACID(Atomicity, Consistency, Isolation, Durability)-compliant transactions across a highly distributed infrastructure

- Latest technologies to appear in the data management area to address Big Data problems
- No proven track record
- Have been adopted by relatively few organizations

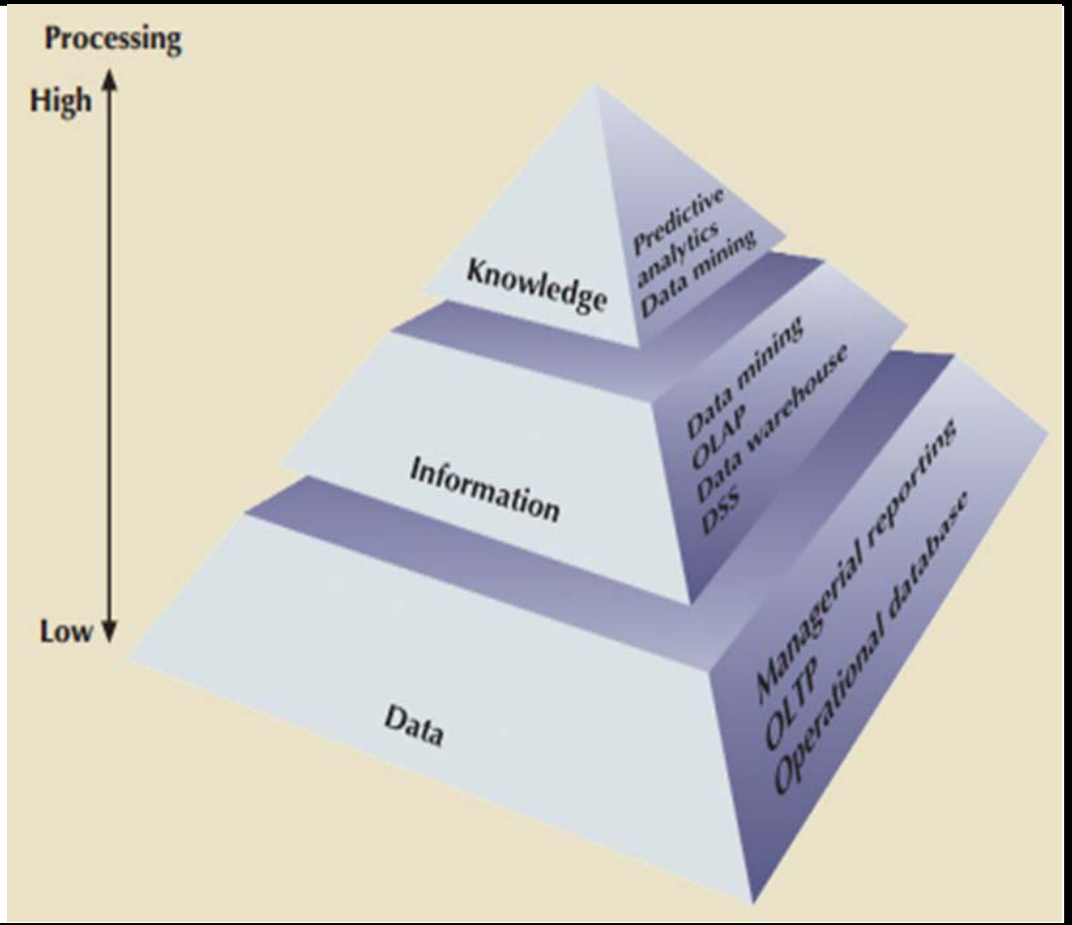
https://www.predictiveanalyticstoday.com/newsql-databases/

DATA ANALYTICS

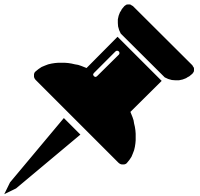



DATA
LOAD
FUTURE
customer
behaviour
BUSINESS
INTELLIGENCE
Useful
DATA MINING
Searching for hidden patterns ...
012012
012012
012012
Decision-making support
This way
Improve
STRATEGY

EXTRACTING KNOWLEDGE FROM DATA



Data Analytics

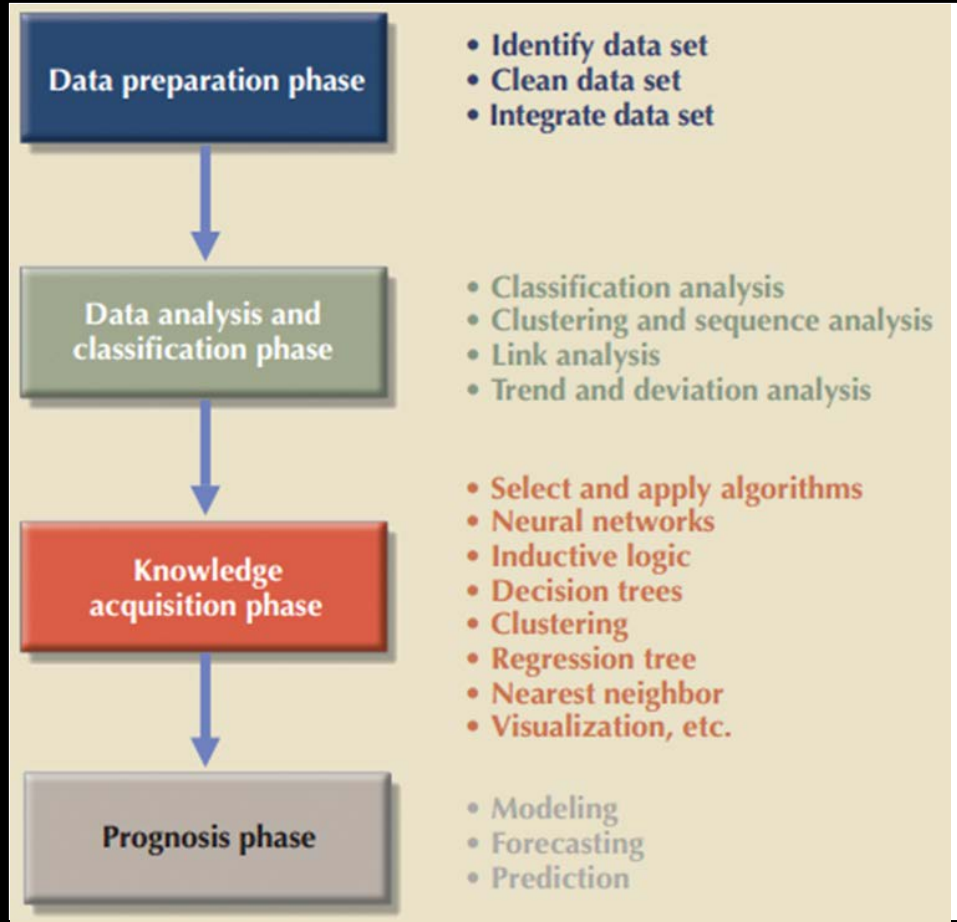


What is Business Intelligence

- The process of collecting raw data or business data and turning it into information that is useful and more meaningful.

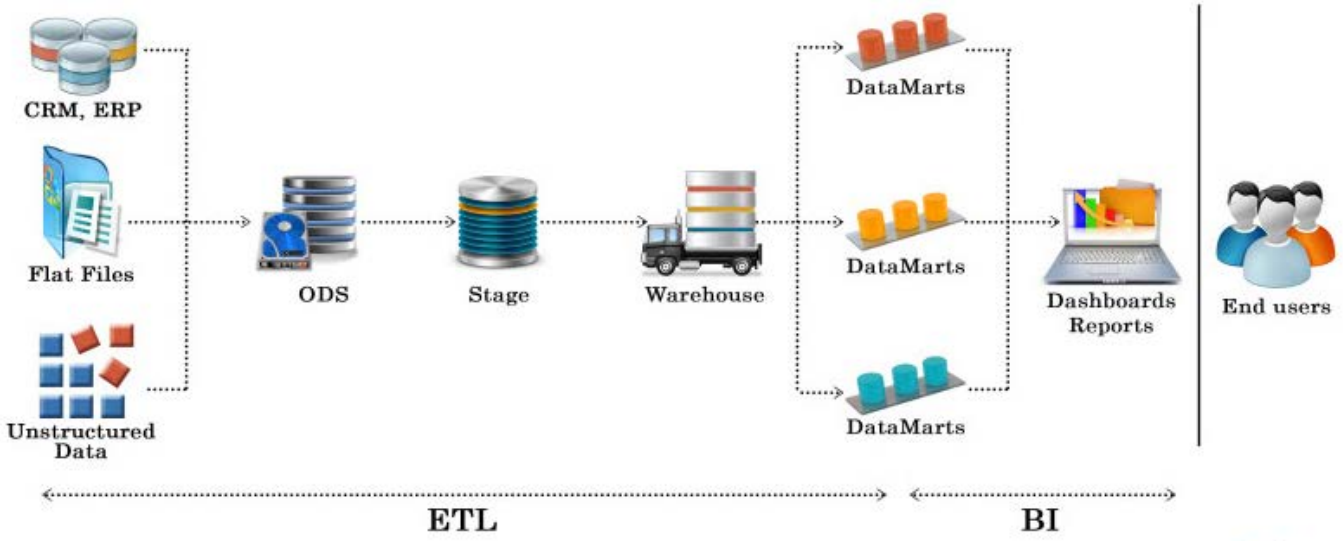
Subset of business intelligence (BI) functionality that encompasses mathematical, statistical, and modeling techniques used to extract knowledge from data

- Continuous spectrum of knowledge acquisition that goes from discovery to explanation to prediction

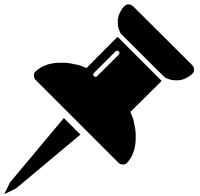


Data- Mining Phases

Data Analytics



Data Analytics Key Terms



Key performance indicator(KPI)

are quantifiable numeric or scale-based measurements that assess the company's effectiveness or success in reaching its strategic and operational goals.

- Ex. Education – Graduation rates, number of incoming students, student retention rates, teaching evaluation scores
- Ex. Human Resources – Applicants to job openings, employee turnover, employee longevity.

Data Mart

A **data mart** is the access layer of the data warehouse environment that is used to get data out to the users. The data mart is a subset of the data warehouse and is usually oriented to a specific business line or team.

Data warehouse

a read-only database optimized for data analysis and query processing

Extract, Transformation and Loading(ETL)

https://en.wikipedia.org/wiki/Data_mart

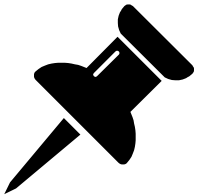
Data Analytics Key Terms

Explanatory Analytics focuses on discovering and explaining data characteristics based on existing data

Predictive Analytics focuses on predicting future data outcomes with a high degree of accuracy

Data Visualization is abstracting data to provide information in a visual format that enhances the user's ability to effectively comprehend the meaning of data.

Predictive Analytics



- Refers to the use of advanced mathematical, statistical, and modeling tools to predict future business outcomes with a high degree of accuracy

- Focuses on creating actionable models to predict future behaviors and events

- Most BI vendors are dropping the term data mining and replacing it with predictive analytics

- Models used in customer service, fraud detection, targeted marketing and optimized pricing

- Can add value in many different ways but needs to be monitored and evaluated to determine return on investment

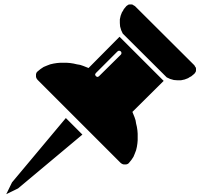
Sample of Business Intelligence Tools

Tool	Description	Sample Vendors
Dashboards and business activity monitoring	Dashboards – use web-based technologies to present key business performance indicators or information in a single integrated view, generally using graphics that are clear, concise, and easy to understand.	Salesforce IBM/Cognos BusinessObjects Information Builders
Portals	Portals – provide a unified, single point of entry for information distributions. Portals are a web-based technology that use a web browser to integrate data from multiple sources into a single webpage. Many different types of BI functionality can be accessed through a portal.	Oracle Portal Actuate Microsoft SAP
Data analysis and reporting tools	These advanced tools are used to query multiple and diverse data sources to create integrated reports.	Microsoft Power BI (https://www.youtube.com/watch?v=Qgam9M8I0xA) MicroStrategy SAS WebReportStudio

Sample of Business Intelligence Tools

Tool	Description	Sample Vendors
Data –Mining tools	These tools provide advanced statistical analysis to uncover problems and opportunities hidden within business data.	SAP Teradata MicroStrategy MS Analytics Services
Data warehouses (DW)	The data warehouse is the foundation of a BI infrastructure. Data is captured from the production system and placed in the DW on a near real-time basis. BI provides company-wide integration of data and the capability to respond to business issues in a timely manner.	Microsoft Oracle IBM/Cognos Teradata
OLAP tools	Online analytical processing provides multidimensional data analysis	IBM/Cognos BusinessObjects Oracle Microsoft
Data Visualization	These tools provide advanced visual analysis and techniques to enhance understanding and create additional insight of business data and it's true meaning	Dundas Tableau QlikView Actuate

JOB OUTLOOK AND GROWTH



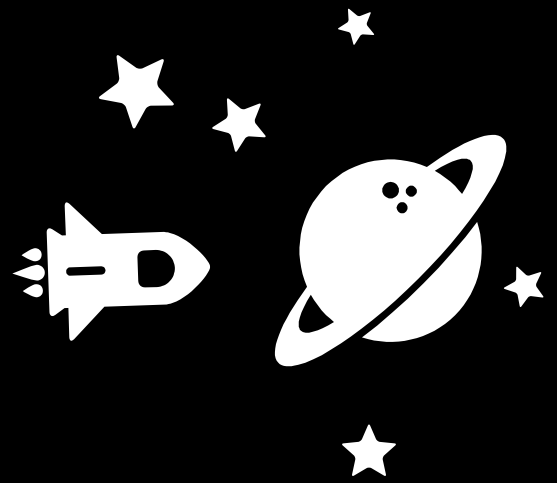
Glassdoor's article on best jobs in America determined by combining three factors: number of job openings, salary, and overall job satisfaction rating.

- https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm

JOB TITLES

- Business Intelligence Analyst
- Data Analyst
- Data Engineer
- Analytics Manager

SAMPLE WEB APP



Building a web application using the MEAN/MERN stack.
MEAN – MongoDB, Express, Angular, NodeJS.
MERN – MongoDB, Express, ReactJS, NodeJS.

TECHNOLOGY STACK

NODEJS
Nodejs.org/en/ - used as the backend processing language. Essentially, it's JavaScript on the server-side that uses an event-driven, non-blocking I/O model that makes it lightweight and efficient

NPM
Npmjs.org - npm is the package manager for JavaScript and the world's largest software registry.

ANGULAR
Angular.io – A frontend javascript framework that is component based.

MONGODB
mongodb.com – The backend database that will use to perform CRUD operations.

MLAB / ROBO 3T
MLAB is a hosting site that host NoSQL/MongoDB databases. They provide you with a free sandbox. Robo 3T is a GUI application that allows you to peer into the database.

EXPRESS
A light framework that allows for routing and building web applications.

STEP 1

DOWNLOAD AND INSTALL NODEJS AND NPM

- Once you have installed NodeJS, open a terminal and run the command: node –version
- You should see the latest version of NodeJS displayed in the terminal, if not make sure NodeJS is correctly installed.
- Also, run the command: npm –version
- You should see the latest version of npm installed, when you installed node, npm also

STEP 2

DOWNLOAD AND INSTALL ANGULAR

- Run command “npm install @angular/cli –g” to install angular
- Run “ng new sampleapp” to set up the project folder structure.
- Change directories and navigate into the project folder “cd sampleapp”
- Run command “ng build” to create a build for our project. We will use express to serve the files from /dist ← was created when you ran the ng build command.
- Running ng build will create a build of our project. We need to do this because our Express server is going to look for a /dist folder to serve the files.

Feel free to refer to Angular quick start: https://angular.io/guide/quickstart

STEP 3

DOWNLOAD AND INSTALL EXPRESS

- Run npm command: npm install express body-parser --save
- Express will serve as middleware for parsing incoming request bodies.
- Create a new file called server.js in the root folder – copy server code from codesnippets.js

STEP 4

DOWNLOAD AND INSTALL MONGODB AND CREATE SANDBOX DATABASE AT MLAB

- Navigate to mlab.com. Create a free sandbox, select US East Region and name the database nflapptest.
- Create a collection called "superbowls" and add three documents for each of the first three superbowls. - JSON objects found in codesnippets.js
- Create a file called api.js and copy and paste api snippet.
- Copy the mongoDB url and input your username and password.
- Replace this url in the MongoClient connection method.
- Run "npm install mongodb –save" – this is a package that will allow you to interact with MongoDB
- Check to make sure database is working and making the api call.
- Run command "node server"
- Go to the browser and type in http://localhost:3000/api/superbowls

STEP 5

CONNECTING ANGULAR AND BACKEND

- Change directories to make sure you are in the app folder - sampleapp/src/app.
- Run command “ng g service data” – this will create a service file called data.service.ts for communicating with the API.
- Open the file and paste the code snippet with for data.service.ts

STEP 6

REGISTERING THE DATA.SERVICE.TS FILE TO ANGULAR

- In Angular, the app.module.ts file act as a resource agent. We have to register everything through this file to use within our application.
- Navigate to your app.module.ts file and copy the code under “Code snippet for app.module.ts” from codesnippet.js and replace it with the code that is currently in app.module.ts.
- There are comments in the code that show you what has been added to the original app.module.ts.

STEP 7

INCORPORATING THE DATA SEVICE FILE INTO OUR COMPONENT

- Now that we have registered the “data.service.ts” file, we can now instantiate an instance of DataService into our app.component.ts file.
- Navigate to your app.componenet.ts file and copy the code under the comment “Implementing app.component.ts” in codesnippet.js and replace it with the code that is currently in app.component.ts.
- There are comments in the code that show you what has been added to the original app.component.ts.

STEP 8

DISPLAYING THE DATA FROM OUR DATABASE

- We have successfully connected our backend to our frontend, all we have to do now is display our data on the front-end.
- Navigate to app.component.html and copy the code under “Implementing app.component.html”
- The code is using a directive to iterate through each object and display to the front-end.

STEP 9

BUILDING OUR PROGRAM AND STARTING THE SERVER

- All of our code is intact, now we need to run the command “ng build” while in the “src/app” folder.
- Once you have ran the “ng build” command, change directories to the root folder by typing “cd ../..”
- Once you are back in the “sampleapp” folder, run the command “node server.js”.
- Open browser at localhost:3000
- Every time you update your application or make a change, you will need to run the “ng build” command from within the “src/app” folder.
- To stop your server using a bash terminal, simply enter “CTRL + C”

STEP 10

INCORPORATING MATERIAL.ANGULAR.IO INTO FRONTEND

- Navigate to “src/app” folder and run the command “npm install --save @angular/material @angular/cdk”
- Also, run the command “npm install --save @angular/animations” to implement animation functionality into the material library.
- By running these two commands, we have simply added these packages to our project, and you can confirm by looking at the package.json file.
- Open your app.module.ts file and copy the updated code under /**Material implementation app.module.ts**/ and paste in the new code.
- Also, navigate to your app.component.ts file and paste in the new code from code snippets titled /**Updated app.component.html implementing Material**/
- After incorporating the code, run the “ng build” command, navigate back to the root folder and run command “node server” to launch server.

Credits

Special thanks to all the people who made and released these awesome resources for free:

- Presentation template by SlidesCarnival