What Is Data Analytics!

Unstructured Data

Unstructured data refers to a dataset that is not stored in a structured format.

Unstructured Data (contd.)

Unstructured data provides a wealth of knowledge that numbers and statistics simply can't explain.

Unstructured Data (contd.)

Unstructured text data goes beyond numerical values and facts into thoughts, opinions, and emotions.
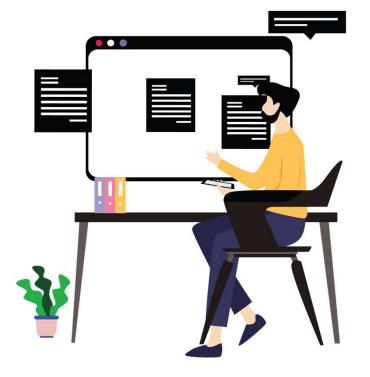
It can be analyzed to provide both quantitative and qualitative results.

Follow market trends, monitor brand reputation, understand the voice of the customer (VoC), and much more.
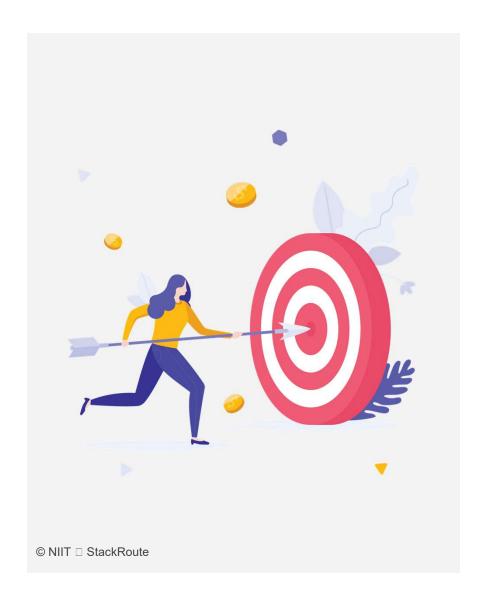
Data Analytics with MongoDB

Learning Objectives

- Distinguish between quantitative and categorical data.

- Explain various analysis techniques.

- Define data exploration with MongoDB.

- Identify data correlation with MongoDB.

- Describe data analysis with embedded document.

News Popularity in Multiple Social Media Platforms

S&P Global Inc. is an American publicly traded corporation headquartered in Manhattan, New York City. Its primary business areas are financial information and analytics.
The company wishes to analyze how a piece of news impacts the business trade.

Comparing Data Types

These are two documents from the news collection.

```
_id: ObjectId("6142f7912226e1e1a22a2df7")
IDLink: "10423"
Title: "A Look at the Health of the Chinese Economy"
Headline: "Tim Haywood, investment director business-unit head for fixed income a..."
Source: "Bloomberg"
Topic: "economy"
PublishDate: "9/20/2008 0:00"
SentimentTitle: "0.208333333"
SentimentHeadline: "-0.156385811"
Facebook: "-1"
GooglePlus: "-1"
LinkedIn: "-1"
```

```
_id: ObjectId("6142f7912226e1e1a22a2df6")
IDLink: "99248"
Title: "Obama Lays Wreath at Arlington National Cemetery"
Headline: "Obama Lays Wreath at Arlington National Cemetery. President Barack Oba..."
Source: "USA TODAY"
Topic: "obama"
PublishDate: "4/2/2002 0:00"
SentimentTitle: "0"
SentimentHeadline: "-0.053300179"
Facebook: "-1"
GooglePlus: "-1"
LinkedIn: "-1"
```
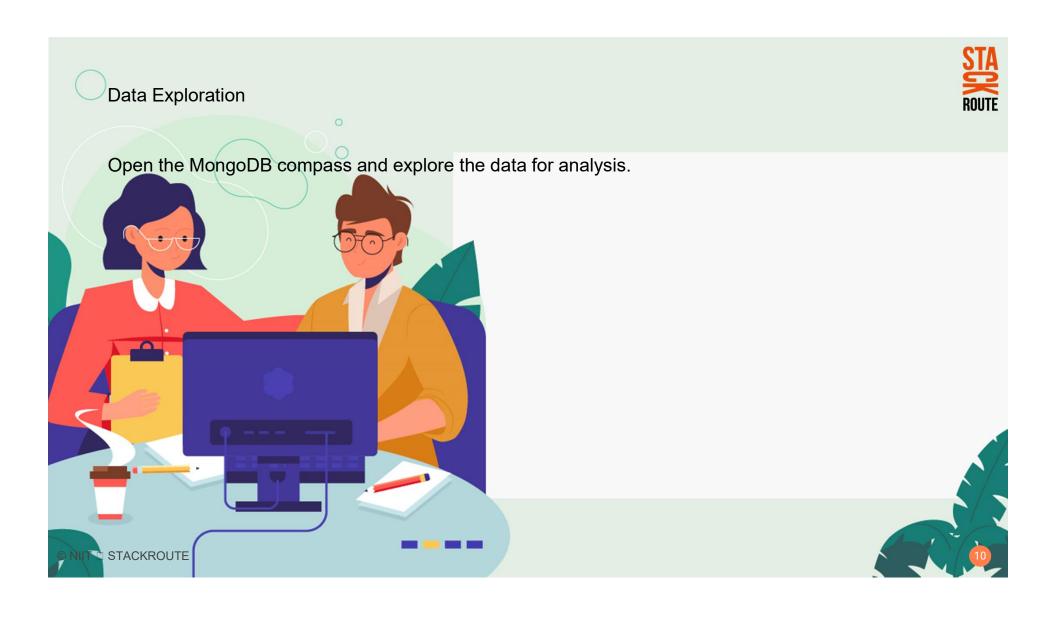
Difference Between Quantitative and Categorical Data

Quantitative data:
- Numerically represented
- Quantified as an amount
- Directly compared with other data

Categorical data:
- Fits into one of the several categories
- Cannot be directly compared with other data

Data Exploration

Open the MongoDB compass and explore the data for analysis.

Sorting and Projection

FILTER { field: 'value' }

PROJECT {topic : 1, Facebook : 1, GooglePlus : 1, LinkedIn: 1, PublishDate: 1}

SORT {GooglePlus : 1}

COLLATION { Locale: 'simple' }

VIEW ≣ {} ▦

_id: ObjectId("6142f7922226e1e1a22a9df0")
PublishDate: "1/19/2016 22:25"
Facebook: "3121"
GooglePlus: "99"
LinkedIn: "357"

_id: ObjectId("6142f7932226e1e1a22ac3bd")
PublishDate: "2/11/2016 21:35"
Facebook: "4726"
GooglePlus: "99"
LinkedIn: "0"

S&P Global Inc. wishes to observe the positive and negative effects on trade based on the popularity of particular news.

Middle Value To Check the Outlier


```
FILTER   {Facebook: {$gte: 24605}}

PROJECT  {Title : 1, Facebook: 1, GooglePlus: 1, LinkedIn: 1}

SORT   {Facebook: -1}

COLLATION   { Locale: 'simple' }
```

```
VIEW   ≣   {}   ▦

_id: ObjectId("61459793f12031e54405eed7")
Title: "Editorial: Welcome rain clouds issues for economy"
Facebook: 49211
GooglePlus: 0
LinkedIn: 0
```

The company also wishes to know about the specific range of popularity values where the trade is starting to increase or decrease.

Up to Date


```
FILTER    {PublishDate: {$gte : new Date("2007-07-02")}}
PROJECT   {Title : 1, Facebook: 1, GooglePlus: 1, LinkedIn: 1, PublishDate : 1 }
SORT      {PublishDate: -1}
COLLATION    { Locale: 'simple' }
```

```
VIEW  ≡  {}  ⊞

_id: ObjectId("61459795f12031e54406acb3")
Title: "Sausalito businesswoman wins $10000 in Microsoft Windows 10 ..."
PublishDate: 2016-07-07T10:46:00.000+00:00
Facebook: -1
GooglePlus: 1
LinkedIn: 0
```

The company also wishes to know the impact of a particular news article on an older date and observe the trend for the current date and also understand how the business trend will be affected on a quarterly basis.

Distinct Values

```
> use News_Article
switched to db News_Article
>

> db.News_popularity.distinct("Topic")
[ "economy", "microsoft", "obama", "palestine" ]
>
```

List out the various trending news for the specific topic.

Filter

The company also search for most trending news topic.

Analyzing the Correlation in Various Fields

The company wishes to know the topic trend on yearly basis.

Analyzing the Correlation in Various Fields (contd.)


```
FILTER  {$and : [{Topic: "microsoft"}, {LinkedIn: {$lt :5000}}]}

PROJECT  {Topic: 1, Facebook: 1, GooglePlus: 1, LinkedIn: 1, PublishDate: 1}

SORT  { field: -1 } or [['field', -1]]

COLLATION  { locale: 'simple' }
```

```
VIEW  :≡  {}  ⊞

_id: ObjectId("61459790f12031e544054082")
Topic: "microsoft"
PublishDate: 2015-02-28T18:49:00.000+00:00
Facebook: -1
GooglePlus: -1
LinkedIn: -1
```

The company is interested in observing the correlation of each topic on the social media.

Analyzing the Correlation in Various Fields (contd.)

The company also takes into consideration the topics on social media.

Analyzing the Correlation in Various Fields (contd.)


```
FILTER  {$or : [{Topic: "palestine"}, {Topic: "obama"}, {Topic: "microsoft"}]}

PROJECT  {topic : 1, Facebook: 1, GooglePlus : 1, LinkedIn: 1, PublishDate: 1}

SORT  { field: -1 } or [['field', -1]]

COLLATION  { locale: 'simple' }

VIEW  ≣  {}  ▦

_id: ObjectId("61459790f12031e54405407d")
PublishDate: 2002-04-01T18:30:00.000+00:00
Facebook: -1
GooglePlus: -1
LinkedIn: -1
```

Which news is trending the most on the current date?

Analyzing the Correlation in Various Fields (contd.)

The company is also interested in observing the impact of the "Microsoft" and "Obama" news on GooglePlus social media site.

Data Analysis with Embedded Documents and Aggregation

Online Exams and Assessments

During the pandemic, it becomes very difficult for the universities to conduct examination and check the students' performance. So, the universities are conducting various types of online assessments including subjective, quiz, and homework assignments.
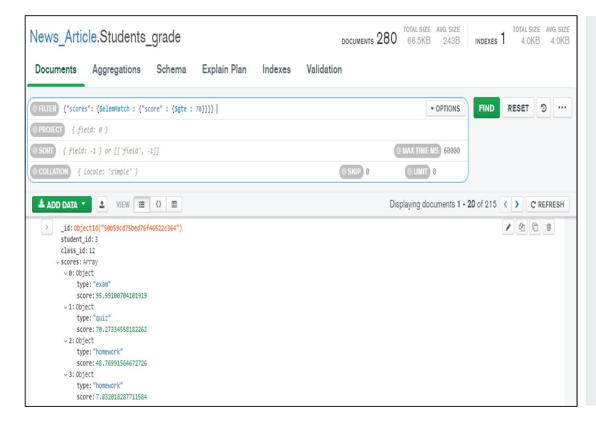
Students need to submit the assignment by specifying the type of examination.
Check whether all the students have submitted the assignments with correct name or not.

Check the number of students who have completed the "homework" assignment.

Identify the students who have a score greater than 70%.

Identify the students who have the exam score less than 70%.

```
1 ▾ /**
2    * _id: The id of the group.
3    * fieldN: The first field name.
4    */
5 ▾ {
6    _id: '$student_id',
7 ▾ max_score: {
8      '$max': '$scores.score'
9    },
10 ▾ min_score:{
11     '$min': '$scores.score'
12   },
13 ▾ count : {
14     '$sum' : 1
15   }
16
17
18 }
```

Output after $group stage ⓘ (Sample of 20 documents)

```
_id: 20
▸ max_score: Array
▸ min_score: Array
  count: 9
```

Group the students' scores based on their id's. Identify the maximum score and the minimum score and count.
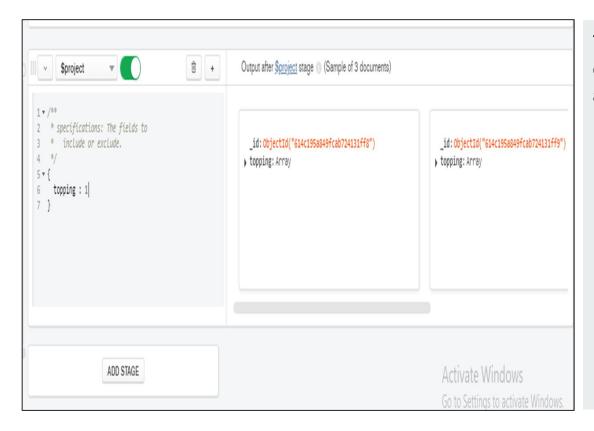
Home-Made-Cakes-and-Donuts

Home-Made-Cakes-and-Donuts is the shop that provides customers choice-based cakes and donuts.

Based on the customer selection of the type of batter and the toppings along with the flavors with various color combinations, the shop prepares cakes and donuts and serves the customer.
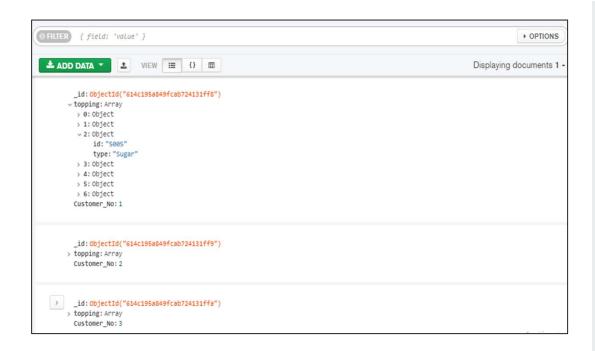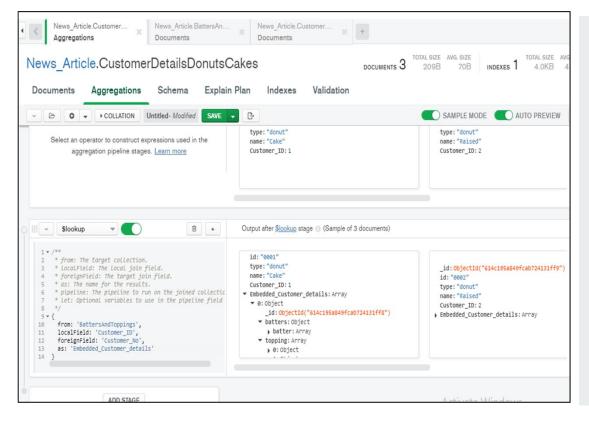
The customer wishes to check the variety of toppings available.

As the customer details are stored in the database, assign the unique identification for each customer for easy access.

Customer wants to check the preferred varieties of toppings.

Thank you!