BIG DATA COURSE

Introduction

First Step towards Data Science!

Introduction

Dimensions of Big Data - Volume, Velocity, and Variety.

Introduction

Other dimensions of Big Data - Veracity, Validity, Volatility, Value, Viability, Visualization, and Variability.

Variability vs. Variety.

Introduction

Forms of Data - Structured, Unstructured, Semi-structured, and Meta-Data.

Introduction

Lifetime access to the content.

Introduction

Data Scientists and Machine Learning Engineers are in demand!

Skills to work with data and obtain insights from it is really important.

Introduction

The goal of this course: Discuss the criteria for Data to be classified as "Big Data"

Introduction

Next Section: Volume.

Volume

Data at rest.

Data which is already present within an enterprise and needs to be processed.

3 Quintillion Bytes of Data is generated every day.

It has become very important to make sense of the data for best business decisions.

Volume

Examples of Volume:

1. Self-Driving Car generating 100's of GBs per year.
2. Social Networks - Facebook, Twitter, Youtube, Instagram, etc
3. Past transactions of a retailer shop or a bank.
4. Clients loyalty cards.
5. GPS devices' data and more.

Volume

40 Zetta Bytes of data will exist by 2020.

90% of all the data ever created, was created in the past 2 years. It is expected to double every year.

Volume

"Volume" is when the size of the data itself becomes a part of the problem.

How much data we have. How rapidly it is growing. This creates distinct storage, management, and processing demands.

Volume

Traditional approaches to managing and process the data fail to complete a task within a certain amount of time or within a certain computational cost.

Relational Databases and Data Warehouses fail to store and retrieve it within a certain amount of time or the computational cost to process it is significant.

Volume

If 10 million rows are "Big" then 9.99 million rows aren't "small".

The decrease in performance indicates:

1. issues of reliability
2. hardware cost
3. long query time

Volume

Inability to handle unstructured or semi-structured data. (discussed in a later section)

System performance decreases as Data Volume increases.

The threshold for maximum performance overhead which is acceptable for a certain business problem.

Volume

In terms of Volume, what is considered 'big' in one enterprise may not be 'big' in another enterprise.

Therefore, the threshold varies from enterprise to enterprise.

Volume

Summary:

1. Volume is the Data which is already present.
2. Size and how quickly it is growing.
3. Data at Rest.
4. When the size of the Data becomes a part of the Problem.
5. No standard, as of the year 2018, for what is considered 'Big' in terms of Volume.

Volume

Next section: Velocity

Velocity

Velocity refers to the data in motion.

How quickly is the data coming in an organization.

How long does it take for an organization to process it.

Velocity implies the speed at which data is being generated.

Velocity

Sources:

Social Networks, Machines, Business Processes, Mobile Devices, etc.

Human Interactions with Machines like GPS, click-stream data, tweets, check-ins. feedbacks, etc.

Continuous generation of large amounts of Data

Velocity

Velocity may be defined as:

The amount of time it takes to process the data and generate insights once it enters their systems or the speed at which data is accessible.

This speed must be in real-time or near real-time.

Velocity

The fast inflow of data requires the enterprises to design highly elastic and available data processing as well as data storage solutions.

Helps businesses and researchers make the best decisions that give them strategic competitive advantages and return over investments.

Velocity

If the results or analysis reports are not generated on time then they are not as valuable.

This is because, for some applications, the data shelf life is really short.

Velocity

Batch algorithms are not suitable for Big Data Solutions.

The algorithms must also be able to produce results in real-time. For example, Stream Computing

Velocity

Examples of Velocity:

1. 3 million Likes per day.
2. 450 million Tweets per day. 300,000 per minute
3. 300 hours of video uploaded to YouTube per minute.
4. 2.5 million Google searches per minute.
5. Personalized Advertisements.
6. Movie Recommendations.
7. Product Recommendations.

Velocity

Summary:

1. Velocity is the speed at which the data is generated, stored, analyzed, and visualized.
2. It refers to the data in motion.

Velocity

Next Section: Variety

Variety

The data in many forms.

Different types of data.

Different sources of data.

Variety

Different types imply that the different formats of data that cannot be stored in traditional, structured, or relational database systems.

Different sources imply that the data is generated and acquired from both inside and outside of the company.

Variety

Presents the biggest challenges of Big Data.

Storing and Retrieving quickly and cost-effectively.

Analyzing all of it together.

The data is rapidly changing.

Variety

Forms of Data:

1. Structured
2. Unstructured
3. Semi-structured
4. Metadata

Variety

Structured Data:

1. Complies with the rules or standards of a data model or a (Relational) schema.
2. Stored in rows and columns form and in a relational database, where it captures the relationships between different entities.
3. Origins: Enterprise Applications and Information Systems.

Variety

Structured Data:

1. Structured Data usually doesn't require any special treatments in terms of Storage and Processing.
2. Examples: inventory Records, Employee Records, Transactions of a Bank, Sale Invoices, Customer Records, etc.

Variety

Unstructured Data:

1. Does not comply with the rules or standards of a data model or a schema.
2. Data is usually textual or binary and is usually in the form of self-contained and non-relational files.

Variety

Unstructured Data:

1. Text Files example: Tweets, Web Content, Books, Emails, Forums, Blog Posts, etc.
2. Binary Files example: Audio, Video, Image, etc.

Variety

Unstructured Data:

1. Makes up 80% of the Data within an Enterprise.
2. Faster growing-rate than Structured Data.
3. Special Processing and Tailored Techniques are required.

Variety

Semi-Structured Data:

1. Some predefined level of structure and consistency.
2. Non-relational in nature.
3. Hierarchical or Graph-Based.

Variety

Semi-Structured Data:

1. Textual in nature.
2. More easily processed and managed than Unstructured Data.
3. Forms of Semi-Structured Data: XML and JSON files.

Variety

Semi-Structured Data:

Spreadsheets, RSS Feeds, Machine-Generated Data (RFID tags), Machine Logs, Cell Phone GPS data, Device Data, and Sensors Generated Data, etc.

Variety

We define a Schema in advance for the Structured Data.

Variety

Metadata:

1. Data about the Data.
2. Provides information about a particular Dataset’s Characteristics and Structure.
3. Machine Generated and Appended to the Original Dataset.

Variety

It provides information about the pedigree of the data as well as its origin and derivation during processing.

Crucial for dealing with Semi-Structured and Unstructured Data.

Variety

Summary:

1. In 2020, 80% of the 40 Zeta Bytes will be unstructured.
2. Analyzing structured data can easily be done using SQL but for unstructured you need to opt for other solutions.

Variety

Summary:

1. Structured data complies with the rules or standards of a data model or a schema.
2. Unstructured data does not comply with the rules or standards of a data model or a schema.
3. Semi-structured data has some predefined level of structure and consistency in it but it is non-relational in nature.
4. Metadata is the data about the data.

Variety

Three V's of Big Data:

1. Volume
2. Velocity
3. Variety

No standard Definition of Big Data.

More V's in the following sections.

Veracity

Data in Doubt.

Veracity refers to the quality or fidelity of the data.

It is the trustworthiness of the data in terms of accuracy.

Veracity is all about making sure that the data is accurate.

Processes to filter out the inaccurate or bad data from entering into the systems.

Veracity

Veracity not only refers to quality and fidelity of the data but it can also be referred to the following:

1. Authenticity
2. Accountability
3. Origin
4. Availability
5. Security
6. Reputation

Veracity

Veracity is the uncertainty about the:

1. Consistency
2. Completeness
3. Latency
4. Deception
5. The model approximations for the data.

Veracity

Inaccurate Input yields inaccurate Output.

"Garbage in, garbage out"

Veracity

Veracity also refers to the biases, noise, and abnormalities in the data being generated.

For meaningful insights, the data must be cleansed first.

Challenging for Big Data.

Veracity

Data enters the systems both from internal and external sources.

The data needs to be assessed for quality as well as requires preprocessing techniques to resolve invalid data and remove noise.

Veracity

The data can either be part of the signal or noise of the dataset.

Noise is the data that cannot be converted into information and thus has no value.

Signals have value and lead to meaningful information.

Veracity

Internal sources or trusted external sources have less noisy data. Tailored acquisition techniques of the data also play a role in its quality.

Uncontrolled acquisition or external untrustworthy sources may have more noise.

The signal-to-noise ratio of the data is dependent upon the source of the data along with the processes used to acquire it.

Veracity

The data is potentially worthless if it is not accurate.

Data accuracy is critical in programs that involve automated decision-making or utilize unsupervised machine learning algorithms.

The results of these programs are as good as the data they are working with.

Veracity

Big Data is noisy and a huge amount of work goes into producing an accurate dataset prior to analysis.

Nearly 70% of the resources are used in Data Cleaning and Preprocessing.

Veracity

Summary:

1. Veracity is also called the Data in Doubt. It refers to the quality of the Data.
2. The accuracy of analysis depends on the veracity of the source data.
3. Veracity begs the question that does the data come from a reliable source and is it accurate and complete?
4. Veracity refers to the biases, noise, and abnormality in data.
5. Having a lot of data in different volumes coming in at high speed is worthless if that data is incorrect.

Veracity

Next section: Validity

Validity

Validity is closely related to the previous dimension, Veracity.

Validity refers to the following with respect to time:

1. Trustworthiness
2. Authenticity
3. Accountability
4. Correctness
5. Appropriateness
6. Precision
7. Accuracy

Validity

How long will the data remain valid or how soon will the data lose its validity?

In some cases, the data will remain valid irrespective of time. But for other cases, the data might be valid for a very short period of time.

Validity

Validity not only deals with the time aspect, but it also refers to the use of correct and accurate data.

The sources for the analysis must be accurate in order to use the results for decision making.

Validity

Summary:

Validity refers to the Trustworthiness, Authenticity, Accountability, Correctness, Appropriateness, Precision, and Accuracy of the data with respect to time.

Validity

Next Section: Volatility

Volatility

Volatility is like a subset of Validity.

It only deals with the time aspect.

Volatility

Volatility refers to how long the data is valid and for how long should it be stored in the systems before it loses its validity.

Data is now being generated in real-time and such algorithms now exist that can leverage this real-time data along with the preexisting data.

Volatility

You need to determine at what point the data is no longer relevant for the current analysis.

Data which is valid right now may or may not be valid after a few days or even after a couple of minutes for the analysis.

Volatility

Next Section: Value

Value

You need to make sure that your organization is getting value from the data.

Value refers to the usefulness of the data for an enterprise or an organization.

Value implies that just having data is of no use unless we can turn it into meaningful insights.

Value

Higher the data accuracy, the more value it holds for the business.

Value is also dependent on how long does the data processing takes because the results of data analytics have a shelf-life.

Value and time are inversely related. The longer it takes for the data to be turned into meaningful insights, the less value it has for a business.

Delayed results hinder the quality and speed of informed decision-making.

Value

The data in itself is not valuable.

The value is in the analysis performed on the data.

The value is in how these organizations use that data.

Value

Value cannot be used to characterize data as "Big Data". All forms of data, both big and small, must provide some sort of value.

Value

Summary:

1. Value refers to the usefulness of the data.
2. Raw data itself is not valuable at all.
3. Value does not characterizes data as "Big Data".

Value

Next Section: Variability

Variability

Variability refers to the data whose meaning is constantly changing.

Variability is different than Variety.

Variability

Coffee Shop example:

1. A few blends of coffee including mocha, cappuccino, and espresso. That's variety.
2. The flavour of the cappuccino blend is not constant over the couple of days you go there. That's variability.

Variability

If the meaning of the data is constantly changing then it can have an impact on the uniformity of the data within an organization.

Variability

Challenges associated with variable data:

1. Consistency:
    a. Is the data consistent?
    b. Inconsistent data hinders the process of data handling, management, and analytics.
2. Accuracy:
    a. Does it accurately portray the events reported?
    b. Important for sentiment analysis.
3. Separability of the signal from noise.
4. Consistent with respect to time.

Variability

You need to understand the context of the data at hand.

Ask the right questions.

Variability

Summary:

1. Variability refers to data whose meaning is constantly changing.
2. Variability is not the same as Variety.
3. Variety is like a lot of flavours. And Variability is the variance in each flavour over time.
4. Variety can have an impact on the uniformity of the data within an organization.
5. Understanding the context of the data can help deal with variability.

Variability

Next Section: Viability

Viability

The ability to work successfully.

Refers to the selection of features.

Refers to the relevance of features according to the business needs and outcomes required.

Also refers to the relationship between the features.

Viability

The goal is to assess the viability of that data.

This is because with so many varieties of data and variables to consider while performing data analytics that we would first want to check a variable's relevance in the analysis.

This will allow us to reduce the cost as well as time in making a fully functional model.

Viability

The secret is to uncover the hidden relationships among these variables or features in order to determine which of these variables are most relevant to the business problem.

Around 5% of the relevant variables or features will get you 95% of the benefit.

This can be done by determining the viability of the selection, relevance, and relationships of these variables or features.

Viability

Examples of viability:

1. Restaurant reviews.
2. Movie reviews.

For Natural Language Processing applications, the top 60-80% of the most frequent words have more viability than the top most frequent words.

Viability

Conclusion:

Viability refers to the ability to work successfully.

In Big Data, Viability refers to the following:

1. The selection of features.
2. The relevance of features with respect to the business problem.
3. The relationship between these features

Through viability, we can determine the most relevant features for a given problem

Viability

Next Section: Visualization

Visualization

Visualization refers to the communication aspect of Big Data.

Visualization is more effective in terms of conveying the meaning or the intended message than using reports containing a stream of numbers, texts, and formulas.

Visualization

Choosing the right analysis and visualization techniques on the data is very challenging.

Telling a complex story using visualization is very difficult but very crucial as well.

Visualization

Visualization can help organizations answer the question they did not know to ask.

It can help them uncover new trends which can be leveraged by the businesses.

Visualization

Summary:

1. Visualization refers to the communication aspect of big data
2. Choosing the right technique for visualization is challenging
3. Visualization can help answer questions and uncover new trends in the data
4. Visualization in Big Data proposes the challenge of dealing with data which is massive, ever-growing, raw, noisy, messy, constantly-changing, and in so many forms.

Conclusion

Congratulations on completing our first course.

Conclusion

Let's go over a quick summary of all the dimensions of Big Data taught in this course

Conclusion

Volume: It refers to the data at rest. It is the data which is already present within an enterprise and needs to be processed.

Conclusion

Velocity: It refers to the data in motion. It is how quickly is the data coming in an organization.

Conclusion

Variety: It refers to the data in many forms. It is the Different types and sources of data.

Conclusion

Veracity: It refers to the quality or fidelity of the data. It is the trustworthiness of the data in terms of accuracy.

Conclusion

Validity: It refers to the trustworthiness and authenticity of the data with respect to time.

Conclusion

Volatility: It refers to how long the data is valid and for how long should it be stored in the systems before it loses its validity.

Conclusion

Value: It refers to the usefulness of the data for an enterprise or an organization.

Conclusion

Variability: It refers to the data whose meaning is constantly changing.

Conclusion

Viability: It refers to the relevance of features and the relationship between the features according to the business needs and outcomes required.

Conclusion

Visualization: It refers to the communication aspect of Big Data.

Conclusion

Good Luck!

BIG DATA COURSE