U20IT701 BIG DATA ANALYTICS
IV YEAR IT - REGULATION 2023
2 MARKS WITH ANSWER
UNIT-I INTRODUCTION TO BIG DATA

1. What is big data approach?

Many It tools are available for big data projects. Organizations whose data workloads are constant and predictable are better served by traditional database whereas organizations challenged by increasing data demands will need to take advantage of Hadoop's scalable infrastructure.

2. List out the applications of big data analytics.

- Marketing
- Finance
- Government
- Healthcare
- Insurance
- Retail

3. List the types of cloud environment.

- Public cloud
- Private cloud

4. What is reporting?

It is the process of organizing data into informational summaries in order to monitor how different areas of a business are performing.

5. What is analysis?

It is the process of exploring data and reports in order to extract meaningful insights which can be used to better understand and improve business performance.

6. List out the cross validation technique.

- Simple cross validation
- Double cross validation
- Multicross validation

7. Write short note on MapReduce?

MapReduce provides a data parallel programming model for clusters of commodity machines. It is pioneered by google which process 20PB of data per day. MapReduce is popularized by Apache Hadoop project and used by Yahoo, Facebook, Amazon and others.

8. What is cloud computing?

Cloud computing is internet-based computing. It relies on sharing computing resources on-demand rather than having local servers or PCS and other devices. It is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort.

9. Describe the drawbacks of cloud computing?

In cloud computing, cheap nodes fail, especially when you have many of them. Mean time between failures(MTBF) for 1 node = 3 years – MTBF for 1000 nodes = 1 day and commodity network has low bandwidth.

10. List out the four major types of resampling.

- Randomized exact test
- Cross-validation
- Jackknife
- Bootstrap
-

11. **What is Big Data?**

Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.

12. List out the best practices of Big Data Analytics.

1. Start at the End
2. Build an Analytical Culture.
3. Re-Engineer Data Systems for Analytics
4 Useful Data Islands.
5. Iterate often.

13. **Write down the characteristics of Big Data Applications.**

a) Data Throttling
b) Computation- restricted throttling
c) Large Data Volumes
d) Significant Data Variety
e) Benefits from Data parallelization

14. **Write down the four computing resources of Big Data Storage.**

a) Processing Capability
b) Memory
c) Storage
d) Network

15. **What is HDFS?**

Apache Hadoop is a collection of open-source software utilities that facilitate using a network of many computers to solve problems involving massive amounts of data and computation. It provides a software framework for distributed storage and processing of big data using the Map Reduce programming model.

16.What is YARN?

YARN is an Apache Hadoop technology and stands for Yet Another Resource Negotiator. YARN is a large-scale, distributed operating system for big data applications. YARN is a software rewrite that is capable of decoupling Map Reduce's resource management and scheduling capabilities from the data processing component.

17. What is Map Reduce Programming Model?

MapReduce is a programming model and an associated implementation for processing and

generating big data sets with a parallel, distributed algorithm on a cluster. The model is a specialization of the split-apply-combine strategy for data analysis.

18. What is Structured Big Data?

Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data. Over the period of time, talent in computer science has achieved greater success in developing techniques for working with such kind of data (where the format is well known in advance) and also deriving value out of it. However, nowadays, we are foreseeing issues when a size of such data grows to a huge extent, typical sizes are being in the rage of multiple zettabytes.

19. What is UnStructured Big Data?

Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it. A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc. Now day organizations have wealth of data available with them but unfortunately, they don't know how to derive value out of it since this data is in its raw form or unstructured format.

20. What is SemiStructured Big Data?

Semi-structured data can contain both the forms of data. We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS. Example of semi-structured data is a data represented in an XML file.

UNIT – II CLUSTERING AND CLASSIFICATION

1. What are the three stages of IDA process?

- Data preparation
- Data mining and rule finding
- Result validation and interpretation
-

2. What is linear regression?

Linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted X. The case of one explanatory variable is called simple **linear regression**.

3. Explain Bayesian Inference ?

Bayesian inference is a method of statistical **inference** in which **Bayes**' theorem is used to update the probability for a hypothesis as more evidence or information becomes available. **Bayesian inference** is an important technique in statistics, and especially in mathematical statistics.

4. What is meant by rule induction?

Rule induction is an area of machine learning in which formal rules are extracted from a set of observations. The rules extracted may represent a full scientific model of the data, or merely represent local patterns in the data.

5. What are the two strategies in Learn-One-Rule Function.

- General to specific
- Specific to general

6. Write down the topologies of Neural Network.

- Single layer

- Multi layer
- Recurrent
- Self-organized
-

7. What is meant by fuzzy logic.

More than data mining tasks such as prediction, classification, etc., fuzzy models can give insight to the underlying system and an be automatically derived from system's dataset. For achieving this, the technique used is grid based rule set.

8. Write short note on fuzzy qualitative modeling.

The fuzzy modeling can be interpreted as a qualitative modeling scheme by which the system behavior is qualitatively described using a natural language. A fuzzy qualitative model is a generalized fuzzy model consisting of linguistic explanations about system behavior in the framework of fuzzy logic instead of mathematical equations with numerical values or conventional logical formula with logical symbols.

9. What are the steps for Bayesian data analysis.

- Setting up the prior distribution
- Setting up the posterior distribution
- Evaluating the fit of the model

10. Write short notes on time series model.

A time series is a sequential set of data points, measured typically at successive times. It is mathematically defined as a set of vectors x(t), t=0,1,2,… where t represents the time elapsed. The Variable x9t0 is treated as a random variable.

11. **Define Clustering.**

Clustering is a popular unsupervised method and an essential tool for Big Data Analysis. Clustering can be used either as a pre-processing step to reduce data dimensionality before running the learning algorithm, or as a statistical tool to discover useful patterns within a dataset.

12. **What are the major clustering methods?**

- Partitioning Method.
- Hierarchical Method.
- Density-based Method.
- Grid-Based Method.
- Model-Based Method.
- Constraint-based Method.

13. **What are the types of clustering?**

- Centroid-based Clustering.
- Density-based Clustering.
- Distribution-based Clustering.
- Hierarchical Clustering.

14. **What is meant by K-means clustering?**

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. ... Data points are clustered based on feature similarity.

15. **Define Decision Trees.**

A Decision Tree is an algorithm used for supervised learning problems such as classification or regression. ... Each leaf of the tree is labeled with a class or a probability distribution over the classes. A tree can be "learned" by splitting the source set into subsets based on an attribute value test.

16. **Define General Algorithm.**

GA's basic working mechanism is as follows: the algorithm is started with a set of solutions (represented by chromosomes) called population. Solutions from one population are taken and used to form a new population (reproduction). This is driven by optimism, that the new population will be superior to the old one.

17. **What is Naive Bayes in Big Data?**

Naive Bayes is a probabilistic technique for constructing classifiers. The characteristic assumption of the naive Bayes classifier is to consider that the value of a particular feature is independent of the value of any other feature, given the class variable.

18. **What is Bayes theorem in Big Data?**

Bayes Theorem is the extension of Conditional probability. Conditional probability helps us to determine the probability of A given B, denoted by P(A|B). So Bayes' theorem says if we know P(A|B) then we can determine P(B|A), given that P(A) and P(B) are known to us.

19. **What are the 4 types of data classification?**

Typically, there are four classifications for data: public, internal-only, confidential, and restricted.

20. **What are the 5 types of data classification?**

- Public data. Public data is important information, though often available material that's freely accessible for people to read, research, review and store
- Private data. ...
- Internal data. ...
- Confidential data. ...
- Restricted data.

UNIT - III ASSOCIATION AND RECOMMENDATION SYSTEM

1. What is data stream model?

A data stream is a real-time, continuous and ordered sequence of items. It is not possible to control the order in which the items arrive, nor it is feasible to locally store a stream in its entirety in any memory device.

2. Define Data Stream Mining.

Data Stream Mining is the process of extracting useful knowledge from continuous, rapid data streams. Many traditional data mining algorithms can be recast to work with larger datasets, but they cannot address the problem of a continuous supply of data.

3. Write short note about sensor networks.

Sensor networks are a huge source of data occurring in streams. They are used in numerous situations that require constant monitoring of several variables, based on which important decisions are made. in many cases, alerts and alarms may be generated as a response to the information received from a series of sensors.

4. what is meant by one-time queries?

One-Time queries are queries that are evaluated once over a point-in-time snapshot of the

data set, with the answer returned to the user.

Eg: A stock price checker may alert the user when a stock price crosses a particular price point.

5. Define biased reservoir sampling.

Biased reservoir sampling is defined as bias function to regulate the sampling from the stream. The bias gives a higher probability of selecting data points from recent parts of the stream as compared to distant past.

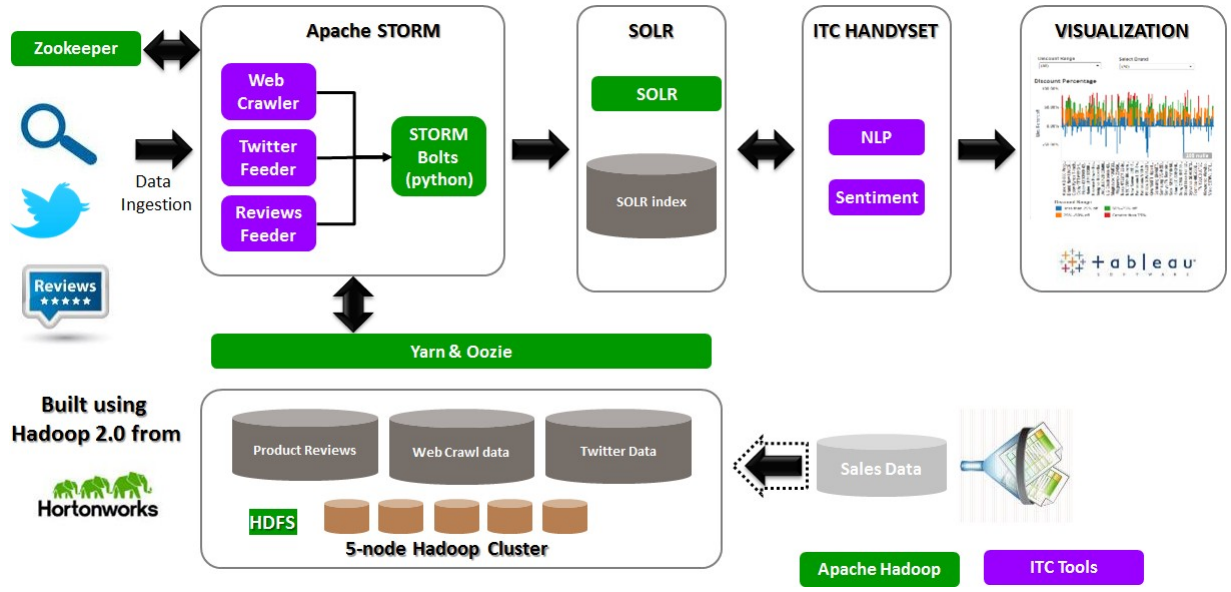
6. What is Bloom Filter?

A Bloom Filter is a space-efficient probabilistic data structure, conceived by Burton Howard Bloom in 1970, that is used to test whether an element is a member of set. False Positive matches are possible but false negative are not, thus a Bloom filter has a 100% recall rate.

7. List out the applications of RTAP.

- Financial services
- Government
- E-Commerce sites

8. Draw a High-Level architecture for RADAR.



9. What are the three layers of Lambda architecture.

- Batch Layer- for batch processing of all data.
- Speed Layer- for real-time processing of streaming data.
- Serving Layer- for responding to queries.

10. What is RTSA?

Real-Time Sentiment analysis (also known as opinion mining) refers to the use of natural language processing text analysis and computational linguistics to identify and extract subjective information in source materials.

11. **What is Association in big data?**

In data science, association rules are used to find correlations and co-occurrences between data sets. They are ideally used to explain patterns in data from seemingly independent information repositories, such as relational databases and transactional databases.

12. **Why is association rule important in big data analysis?**

These large data need to be analyzed in order to extract useful knowledge and present it to decision makers for further use. ... The term of association rules is a powerful technique of data mining for discovering correlation and relationships between objects in the database.

13. **What are the applications of association rule?**

Applications of association rule mining are stock analysis, web log mining, medical diagnosis, customer market analysis bioinformatics etc. In past, many algorithms were developed by researchers for Boolean and Fuzzy association rule mining such as Apriori, FP-tree, Fuzzy FP-tree etc

14. **Why is the association rule especially important in big data analysis?**

This technique is particularly appropriate for analyzing the correlations between objects, because it considers conditional interaction among input data sets, and produce the decision rules of the form IF-THEN. ... Since the datasets are extremely large, parallel algorithms are required.

15. **What is the application of Apriori algorithm?**

Apriori algorithm is a classical algorithm in data mining. It is used for mining frequent itemsets and relevant association rules. It is devised to operate on a database containing a lot of transactions, for instance, items brought by customers in a store.

16. **Define Apriori Algorithm.**

Apriori is an algorithm for frequent item set mining and association rule learning over relational databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database.

17. **What are the basic steps in the Apriori algorithm?**

- Computing the support for each individual item. The algorithm is based on the notion of support.
- Deciding on the support threshold. ...
- Selecting the frequent items. ...
- Finding the support of the frequent itemsets. ...
- Repeat for larger sets. ...
- Generate Association Rules and compute confidence. ...
- Compute lift.

18. **What are two steps of Apriori algorithm?**

Apriori algorithm was the first algorithm that was proposed for frequent itemset mining. It was later improved by R Agarwal and R Srikant and came to be known as Apriori. This algorithm uses two steps “join” and “prune” to reduce the search space. It is an iterative approach to discover the most frequent item sets.

19. **Define Collaborative Recommendation in big data?**

There are two types of recommendations – User Based and Item Based Collaborative Filtering. In User Based Collaborative Filtering, to recommend products for a given user, we compute similarity between the user and every other user in the site. Similarity is computed using distance algorithms or correlation algorithms.

20. **What is Knowledge Based Recommendation in Big Data?**

Knowledge-based recommender systems (knowledge based recommenders) are a specific type of recommender system that are based on explicit knowledge about the item assortment, user preferences, and recommendation criteria (i.e., which item should be recommended in which context).

UNIT-IV STREAM MEMORY

1. What is Association Rule Mining?

The Association Rule Mining is main purpose to discovering frequent itemsets from a large dataset is to discover a set of if-then rules called Association rules. The form of an association rules is $I \rightarrow j$, where I is a set of items(products) and j is a particular item.

2. List any two algorithms for Finding Frequent Itemset.

- Apriori Algorithm
- FP-Growth Algorithm
- SON algorithm
- PCY algorithm

3. What is meant by curse of dimensionality?

Points in high-dimensional Euclidean spaces, as well as points in non-Euclidean spaces often behave unintuitively. Two unexpected properties of these spaces are that the random points are almost always at about the same distance, and random vectors are almost always orthogonal.

4. Write an algorithm of Park-Chen-Yu.

```
FOR(each basket):
    FOR(each item in basket):
        add 1 to item's count;
    FOR(each pair of items):
        {hash the pair to a bucket;
         add 1 to the count for that bucket:}
```

5. Define Toivonen's Algorithm

Toivonen's algorithm makes only one full pass over the database. The algorithm thus produces exact association rules in one full pass over the database. The algorithm will give neither false negatives nor positives, but there is a small yet non-zero probability that it will fail to produce any answer at all. Toivonen's algorithm begins by selecting a small sample of the input dataset and finding from it the candidate frequent itemsets.

6. List out some applications of clustering.

- Collaborative filtering
- Customer segmentation
- Data summarization
- Dynamic trend detection
- Multimedia data analysis
- Biological data analysis
- Social network analysis

7. What are the types of Hierarchical Clustering Methods.

- Single-link clustering

- Complete-link clustering
- Average-link clustering
- Centroid link clustering

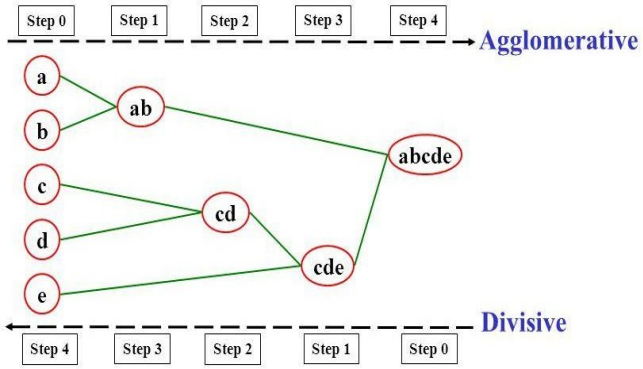
8. Define CLIQUE

CLIQUE is a subspace clustering algorithm that automatically finds subspaces with high-density clustering in high dimensional attribute spaces. CLIQUE is a simple grid-based method for finding density-based clusters in subspaces. The procedure for this grid-baased clustering is relatively simple.

9. What is meant by k-means algorithm?

The family of algorithms is of the point-assignment type and assumes a Euclidean space. It is assumed that there are exactly *k* clusters for some known *k*. After picking *k* initial cluster centroids, the points are considered one at a time and assigned to the closest centroid.

10. Draw the diagram for Hierarchical Clustering.



11. Define Stream Memory.

Streaming data is an analytic computing platform that is focused on speed. Therefore, data is continuously analyzed and transformed in memory before it is stored on a disk. Processing streams of data works by processing "time windows" of data in memory across a cluster of servers.

12. What is in memory streaming?

With in-memory stream processing platforms, you can respond to data on-the-fly, prior to its storage, enabling ultra-fast applications that process new data at the speed with which it is generated.

13. **What is Filtering Streams in Big Data?**

Filtering condition of a stream item is independent of other items of the same stream or any other data stream. The most common example of such filtering is stream sampling, when each item is filtered out with a certain probability and the remaining items form the desired sample.

14. **What are the different types of data streams?**

- Sensor readings from machines.
- e-Commerce purchase data.
- Stock exchange data to predict the stock price.
- Credit card transactions for fraud detection.
- Social media sentiment analysis.

15. **What is a data stream example?**

Dynamic data that is generated continuously from a variety of sources is considered streaming data. ... Log files, e-commerce purchases, weather events, utility service usage, geo-location of people and things, server activity, and more are all examples where real-time streaming data is created.

16. **What is real time sentiment analysis in big data?**

Real-time sentiment analysis is an AI-powered solution to track mentions of your brand and products, wherever they may appear, and automatically analyze them with almost no human input needed.

17. **What is time sentiment analysis?**

Sentiment Analysis is a Natural Language Processing technique to predict the sentiment or opinion of a given text. It involves the use of NLP, text analysis, computer linguistic to identify or extract subjective information.

18. **What is graph analytics in big data?**

Graph analytics uses algorithms to explore the relationships among entries in a graph database, including connections among different people, transactions or organizations. Use cases include contact tracing, cybersecurity, drug interaction, recommendation engines, social networks and supply chains.

19. **What is a decaying window?**

Considering the example of finding the most popular movies in an stream of ticket sales: We shall use an exponentially decaying window with a constant c, which you might think of as 10−9. That is, we approximate a sliding window holding the last one billion ticket sales. Suppose the new ticket is for movie M.

20. **How is data analysis used in stock market predictions?**

A time series model is created by using machine learning and/or deep learning models to accumulate the price data. The data needs to be analyzed and then fitted to match the model. This is what makes it possible to predict future stock prices over a set timetable.

UNIT-V NOSQL DATA MANAGEMENT FOR BIG DATA AND VISUALIZATION

1. What are the main goals of Hadoop?

- Saclable
- Fault tolerance
- Economical
- Handle hardware failures.

2. What is hive?

Hive provides a warehouse structure for other Hadoop input sources and SQL-Like access for data in HDFS. Hive's query language, HiveQL, compiles to MapReduce and also allows user-defined functions(UDFS).

3. What are the responsibilities of MapReduce Framework?

- Provides overall coordination of execution.
- Selects nodes for running mappers.
- Starts and monitors mapper's execution.
- Sorts and shuffles output of mappers.
- Chooses locations for reducer's execution.
- Delivers the output of mapper to reducers node.
- Starts and monitors reducers's execution.

4. What is a Key-Value store?

The key-value store uses a key to access a value. The key-value store has a schema-less format. The key can be artificially generated or auto-generated while the value can be string, JSON, BLOB, etc. the key-value uses a hash table with a unique key and a pointer to a particular item of data.

5. What is visualization? What are the three major goals in visualization.

Visual Visualization is the presentation or communication of data using interactive interfaces. It has three major goals:

- Communicating/presenting the analysis results efficiently and effectively.
- As a tool for confirmatory analysis that is to examine the hypothesis, analyze and confirm.
- Exploratory data analysis as an interactive and mostly undirected search for finding structures and trends.

6. What is sharding?

Horizontal partitioning of a large database leads to partitioning of rows of the database. Each partition forms part of a shard, meaning small part of the whole. Each part can be located on a separate database server or any physical location.

7. **Define NoSQL Databases in big data.**

NoSQL Database is a non-relational Data Management System, that does not require a fixed schema. It avoids joins, and is easy to scale. The major purpose of using a NoSQL database is for distributed data stores with humongous data storage needs.

8. **How NoSQL is used in big data analytics?**

NoSQL allows for high-performance, agile processing of information at massive scale. It stores unstructured data across multiple processing nodes, as well as across multiple servers. As such, the NoSQL distributed database infrastructure has been the solution of choice for some of the largest data warehouses.

9. **What are the 4 types of NoSQL databases?**

In crux, we can say that there are four types of NoSQL Databases: Key-Value (KV) Stores, Document Stores, Column Family Data stores, and Graph Databases.

10. **Which applications use NoSQL?**

Since NoSQL database store the data in schema-less for the application developer can update the apps without having to do major modification in database. The mobile app companies like Kobo and Playtika, uses NOSQL and serving millions of users across the world.

11. **What is the use of NoSQL?**

NoSQL Database is a non-relational Data Management System, that does not require a fixed schema. It avoids joins, and is easy to scale. The major purpose of using a NoSQL database is for distributed data stores with humongous data storage needs. NoSQL is used for Big data and real-time web apps.

12. **What is the difference between NoSQL and MySQL?**

MySQL is a relational database that is based on tabular design whereas NoSQL is non-relational in nature with its document-based design. ... MySQL is one of the types of relational database whereas NoSQL is more of design based database type with examples like MongoDB, Couch DB, etc.

13. **How is big data used in e-commerce?**

Big data allows e-commerce businesses to understand customers better through customer behavior analysis. Big data resources enable optimization logistics, supply chain management, and operational process. This contributes to better performance and significant cost reductions.

14. **What is the role of big data analytics in e-commerce?**

Big Data Analytics (BDA) aims to improve the decision-making process by analyzing and understanding big data, e.g., messages, social media posts, etc. Furthermore, BDA capabilities are used in e-commerce activities as a key growth direction to increase vendors' revenues and attract customers.

15. **What is graph database in big data?**

A graph database is defined as a specialized, single-purpose platform for creating and manipulating graphs. Graphs contain nodes, edges, and properties, all of which are used to represent and store data in a way that relational databases are not equipped to do.

16. **What is graph database with example?**

With the Graph Database model, Digital Asset Management becomes intuitive. Graph Database Example: Netflix uses Graph Database for its Digital Asset Management because it is a perfect way to track which movies (assets) each viewer has already watched, and which movies they are allowed to watch (access management).

17. Which is the best graph database?

Neo4j has the most popular and active graph database community. Reviews report that their product is easy to learn and easy to use with plenty of resources from training materials to books. Neo4j is well-established with loads of resources for their users.

18. How does a graph database store data?

Graph data is kept in store files, each of which contain data for a specific part of the graph, such as nodes, relationships, labels and properties. Dividing the storage in this way facilitates highly performant graph traversals

19. How does Twitter use big data?

The main motivation for the Twitter trend analysis is to **identify the recent trends happening across the world** using big data machine learning techniques. This will help to analyze what has happened in the past and what may happen in the future. ... Twitter API provides a standard way to read and write Twitter data.

20. What big data does Twitter generate?

Using Twitter's API, you get back what's called a JSON output that includes a large amount of information on the user. That JSON output is shown to the right.