Which Hadoop processing model is best suited for batch processing of large datasets?
{ ~Spark Streaming ~Pig ~Real-time processing =MapReduce }

Which phase in the MapReduce framework is responsible for aggregating and summarizing the results fro m Mapper tasks?
{ ~ Map phase ~Shuffle and Sort phase =Reduce phase ~Finalization phase}

What is the primary function of the Mapper in the Hadoop MapReduce framework?
{ ~Aggregating data ~Sorting data ~Combining data =Transforming and extracting data }

In the MapReduce paradigm, what is the role of the Reducer?
{~Splitting input data into smaller chunks =Merging intermediate key-value pairs ~Mapping data to appro priate keys ~Distributing data across the cluster}

Which phase of the MapReduce process involves grouping and shuffling intermediate key-value pairs bas ed on keys before sending them to reducers?
{ ~Mapping phase ~Sorting phase ~Reducing phase =Shuffling phase }

What is the output format of the Mapper in MapReduce?
{ =Key-value pairs ~Sorted data ~Aggregated values ~Raw input data }

What is the primary purpose of the Reducer in the MapReduce framework?
{ ~Performing data transformations ~Filtering out unnecessary data ~Distributing tasks across nodes =A ggregating and summarizing data }

Which phase of the MapReduce process involves executing the user-defined logic to process and analyz e the data?
{ ~Mapping phase ~Shuffling phase =Reducing phase ~Sorting phase }

What is the typical input to the Mapper in the MapReduce process?
{ =Raw data from HDFS ~Sorted data from previous jobs ~Intermediate key-value pairs ~Aggregated dat a }

In the MapReduce paradigm, what is the key role of the Reducer in the final output?
{ ~Grouping and shuffling data ~Splitting data into smaller chunks ~Sorting data =Generating the final out put }

What are the primary phases of the MapReduce framework?
{ ~Load and Store =Map and Reduce ~Shuffle and Sort ~Merge and Combine }

Which phase of the MapReduce framework is responsible for breaking down the input data into smaller c hunks and distributing them to worker nodes?
{ =Splitting Phase ~Mapping Phase ~Shuffling Phase ~Reducing Phase }

What is the purpose of the Shuffling phase in MapReduce?
{ ~It sorts the intermediate key-value pairs. ~It aggregates the final results. =It transfers intermediate data between Map and Reduce tasks. ~It generates the initial key-value pairs. }

Which phase(s) involve(s) parallel processing of data in the MapReduce framework?
{ ~Mapping ~Shuffling ~Reducing =Both Mapping and Reducing }

Which of the following is a key feature of the MapReduce framework that enables fault tolerance?
{~Automatic load balancing ~Data shuffling =Data replication ~Combining phase}

What is the final output of the MapReduce job?

{ ~Intermediate key-value pairs ~Partially processed data =Final key-value pairs ~Raw input data }

Which phase ensures that all values associated with a single key are brought together before the Reducin g phase?
{ ~Mapping =Shuffling ~Combing ~Reducing }

In which phase is the user-defined Reducer function applied to the grouped intermediate data?
{ ~Mapping ~Shuffling =Reducing ~Combining }

What role does the Combiner function play in the MapReduce framework?
{ ~It combines multiple input files into a single file. ~It performs an initial processing step before the Mapp ing phase. =It optimizes the data transfer between Map and Reduce tasks.
~It generates intermediate key-value pairs. }

 In the MapReduce model, what role do worker nodes play?
{ ~They are responsible for distributing tasks to the master node. =They process input data in parallel. ~T hey execute the reducer functions. ~They handle data shuffling and sorting.}

What is the primary advantage of using the MapReduce model for processing large datasets?
{ ~Real-time processing of data ~Elimination of data duplication ~Simplified programming model for all ty pes of applications =Scalability for parallel and distributed processing }

Which company initially developed the MapReduce programming model and framework?
{ =Google ~Microsoft ~Amazon ~IBM }

Which programming language is commonly used for writing MapReduce jobs in Hadoop?
{ ~Python =Java ~C++ ~Ruby }

the "Counting" design pattern in MapReduce is often used for:
{ ~Counting the number of Mapper tasks executed. ~Counting the number of keys in the output. =Countin g the occurrences of specific values in the dataset. ~Counting the total number of Map and Reduce phase s. }
During the shuffling phase, what is the key criteria for grouping intermediate key-value pairs?
{ ~Value content ~Partitioning logic =Key content ~Mapper node ID }

The world’s largest Hadoop cluster?
{ ~Apple =Facebook ~Datamatics ~None of the mentioned }

These are the given are completely describe Hadoop, EXCEPT?
{ ~Open-source =Real-time ~Distributed computing approach ~Java-based }

The client reading the data from HDFS filesystem in Hadoop does which of the following?
{ ~Gets only the block locations form the namenode ~Gets the data from the namenode =Gets both the data and block location from the namenode ~Gets the block location from the datanode }

Amongst which of the following represents the Use of Hadoop
{ ~Robust and Scalable ~Affordable and Cost Effective ~Adaptive and Flexible =All of the mentioned ab ove }

Point out the worng statement?
{=Non relational databases require that schemas be defined before you can add data ~No sql database a re built to allow the insertion of data without a predefined schema ~Both ~None }

These are the given are completely describe Hadoop, EXCEPT?
{ ~Open-source =Real-time ~Distributed computing approach ~Java-based }

What was Hadoop written in?
{ ~Java (software platform) ~Perl =Java (programming language) ~Lua (programming language) }

which is the slave/worker node and holds the user data in the form of Data Blocks?
{ =DataNode ~NameNode ~Data block ~Replication }

As compared to RDBMS, Hadoop?
{ ~Has higher data Integrity ~Does ACID transactions ~Is suitable for read and write many times =Works better on unstructured and semi-structured data}

For YARN, the Manager UI provides host and port information?
{~Data Node ~NameNode =Resource ~Replication }

Hadoop is a framework that allows the distributed processing of?
{ ~Small Data Sets ~Semi-Large Data Sets =Large Data Sets ~Large and Small Data sets }

Hadoop is a framework that works with a variety of related tools Common cohorts include?
{ =MapReduce Hive and HBase ~MapReduce MySQL and Google Apps ~MapReduce Hummer and Igua na ~MapReduce Heron and Trumpet }

Hadoop is open source?
{ ~Always True =True only for Apache Hadoop ~True only for Apache and Cloudera Hadoop ~Always Fal se }

Hadoop run Which of the following platforms?
{ =Cross-platform ~Debian ~Bare-metal ~Unix-like }

Point out the correct statement?
{ =Hive is not a relational database, but a query engine that supports the parts of SQL specific to querying data ~Hive is a relational database with SQL support ~Pig is a relational database with SQL support ~All of the mentioned }

The client reading the data from HDFS filesystem in Hadoop does which of the following?
{ ~Gets only the block locations form the namenode ~Gets the data from the namenode =Gets both the d

ata and block location from the namenode ~Gets the block location from the datanode }

What was hadoop named after?
{ ~Creator Doug cutting's favorite circus act ~Cutting's high school rock band =The toy elephant of cutting's son ~none of the above }

Which one of the following stores data?
{ ~Name node =Datanode ~Master node ~None of these }

Which one of the following stores Metadata?
{ =Name node ~Datanode ~Master node ~None of these }

What is the minimum amount of data that a disk can read or write in HDFS?
{ ~Byte size =Block size ~Heap ~None of the above }

A _____ serves as the master and there is only one NameNode per cluster.
{ ~Data Node =NameNode ~Data block ~Replication }

_____ NameNode is used when the Primary NameNode goes down.
{ ~Rack ~Data =Secondary ~None of the mentioned }

Point out the wrong statement.
{ ~Replication Factor can be configured at a cluster level (Default is set to 3) and also at a file level ~Block Report from each DataNode contains a list of all the blocks that are stored on that DataNode ~User data is stored on the local file system of DataNodes =DataNode is aware of the files to which the blocks stored on it belong to }

The need for data replication can arise in various scenarios like _____
{ ~Replication Factor is changed ~DataNode goes down ~Data Blocks get corrupted =All of the mentioned }

For YARN, the _____ Manager UI provides host and port information.
{ ~Data Node ~NameNode =Resource ~Replication }

During start up, the _____ loads the file system state from the fsimage and the edits log file.
{ ~DataNode =NameNode ~ActionNode ~None of the mentioned }

_____ is a platform for constructing data flows for extract, transform, and load (ETL) processing and analysis of large datasets.
{~Pig Latin ~Oozie =Pig ~Hive }

_____ is the architectural center of Hadoop that allows multiple data processing engines.
{ =YARN ~Hive ~PIG ~Incubator }

The _____ is a framework-specific entity that negotiates resources from the ResourceManager.
{ ~NodeManager ~ResourceManager =ApplicationMaster ~All of the mentioned}

Point out the correct statement.
{ ~YARN also extends the power of Hadoop to incumbent and new technologies found within the data center ~YARN is the central point of investment for Hortonworks within the Apache community ~YARN enhances a Hadoop compute cluster in many ways =All of the mentioned }

Which Hadoop ecosystem tool is used for querying and managing large datasets stored in HDFS using SQL-like queries?
{ ~Pig ~Flume ~Sqoop =Hive }

Which Hadoop component is used for processing and analyzing large datasets in parallel across a cluster ?
{ ~HBase ~Oozie =Spark ~Flume }

What is Hadoop?
{ ~A programming language ~A data storage system ~An operating system =A distributed computing fra mework }

Which Hadoop component is a NoSQL database that provides real-time read/write access to large datase ts?
{ =HBase ~Hive ~Oozie ~Kafka }

Which Hadoop ecosystem tool is used for ingesting and collecting streaming data from various sources?
{ =Flume ~Oozie ~Hue ~Sqoop }

Which Hadoop ecosystem component is used for workflow scheduling and coordination of Hadoop jobs?
{ ~HBase =Oozie ~Sqoop ~Spark }

Which component of Hadoop provides fault tolerance by replicating data across nodes in the cluster?
{ ~YARN (Yet Another Resource Negotiator) =HDFS (Hadoop Distributed File System) ~Spark ~Hive }

Which technology is commonly associated with ACID (Atomicity, Consistency, Isolation, Durability) proper ties for ensuring data integrity?
{ ~Hadoop ~NoSQL databases =RDBMS ~Apache Spark }

Which data model is typically used by RDBMS?
{ ~Key-Value ~Document ~Columnar =Tabular }

In which industry is Hadoop commonly used to analyze and process large volumes of customer data for in sights and personalized recommendations?
{ ~Agriculture ~Construction ~Healthcare =E-commerce }

Which use case involves using Hadoop to process and analyze sensor data from various devices to monit or and optimize industrial processes?
{ ~Social media analysis ~Fraud detection =Internet of Things (IoT) applications ~Financial risk assess ment }

In HDFS, what is the default block size used for splitting files into data blocks?
{ ~64 MB =128 MB ~256 MB ~512 MB }

Which component of HDFS is responsible for managing metadata, such as namespace information and fil e-to-block mapping?
{ ~DataNodes ~ResourceManager =NameNode ~SecondaryNameNode }

What is the function of DataNodes in HDFS architecture?
{ ~Manage metadata ~Execute MapReduce tasks =Store and manage actual data blocks ~Schedule an d allocate cluster resources }

Which HDFS operation is responsible for copying data blocks from one node to another to maintain data r eplication and fault tolerance?
{ ~Data replication ~Block migration =Block replication ~Block balancing }

In HDFS, how does the system ensure data reliability and fault tolerance?
{ =By storing multiple copies of data blocks on different nodes ~By using RAID (Redundant Array of Inde

pendent Disks) for data protection ~By encrypting the data blocks~By compressing the data blocks }

Which feature of Hadoop allows it to scale horizontally by adding more nodes to the cluster as needed?
{ ~Vertical scalability =Elasticity ~Replication factor ~Sharding }

What is the purpose of a NameNode in HDFS?
{ ~Storing data blocks ~Executing MapReduce jobs =Managing metadata and namespace information ~Coordinating resource allocation }

What is the primary advantage of using Hadoop for data processing?
{ ~Real-time processing of data ~Centralized data storage ~Strong schema enforcement =Scalability for processing large datasets }

Which of the following is a characteristic of NoSQL databases?
{ ~They only support structured data ~They are limited to relational data models =They are designed for horizontal scalability ~They use only the SQL query language }

In NoSQL databases, what is the CAP theorem concerned with?
{ ~Query performance ~Data encryption ~Data modeling =Consistency, Availability, and Partition tolerance }

Which NoSQL database type is suitable for storing and managing semi-structured data, such as JSON or XML documents?
{ ~Key-Value ~Columnar =Document ~Graph }

Which of the following is an example of a popular NoSQL database management system?
{ ~MySQL ~Oracle Database =MongoDB ~PostgreSQL }

What challenge in distributed computing involves designing systems that can handle an increasing number of users, requests, or data while maintaining performance?
{ ~Load balancing =Scalability ~Fault tolerance ~Data consistency }

Which challenge pertains to the issue of data synchronization and ensuring that updates from one node are properly propagated to other nodes in a distributed system?
{ ~Concurrency ~Fault tolerance ~Data integrity =Consistency }

What is the term for information, facts, or statistics that are collected and organized for analysis?
{ ~Metadata =Data ~Bigdata ~Analytics }

Which characteristic of big data refers to the rapid generation of data from various sources?
{ =Velocity ~Variety ~Volume ~Validity }

Which type of data lacks a specific structure and includes text, images, and videos?
{ ~Structured data ~Numerical data =Unstructured data ~Semistructured data }

Social media posts and emails are examples of which type of digital data?
{ =Text data ~Image data ~Audio data ~Video data }

Which of the following is a traditional source of data?
{ ~Social media ~Sensors =Surveys ~IoT devices }

How does natural language processing (NLP) assist in working with unstructured data?
{ ~It organizes data into a structured format ~It helps analyze structured data =It processes and understa nds human language in text ~It converts unstructured data into numbers }

Big data is characterized by three Vs. Which of the following is NOT one of the three Vs?
{ =validity ~Variety ~Volume ~Velocity }

What is the primary need for big data in various industries?
{ ~To increase data storage costs ~To slow down decision-making processes =To uncover valuable insig hts and make informed decisions ~To reduce the need for data analysis }

Which category would you place the consumer complaints and feedback?
{ ~structured data =unstructured data ~Numerical data ~Semistructured data }

What category will you place CCTV footage into?
{ ~structured data =unstructured data ~Numerical data ~Semistructured data }

Business Intelligence (BI) involves
{ ~Collecting and storing massive amounts of data ~Making decisions based solely on intuition =Analyzi ng data to make informed decisions ~Conducting surveys to understand customer preferences }

Data science involves
{ ~Generating random data for analysis ~Studying data from a single source =Extracting insights from da ta using various techniques ~Creating complex algorithms without using data }

What is the primary goal of big data analytics?
{ ~Storing as much data as possible ~Summarizing historical data =Extracting meaningful insights from l arge datasets ~Automating data entry processes }

What is the importance of big data analytics?
{ ~It increases data storage costs ~It replaces human decision-making =It helps organizations gain insig hts and make informed decisions ~It eliminates the need for data collection }

What is a major challenge in big data analytics?
{ ~Lack of data privacy concerns ~Limited variety of data sources =Difficulty in extracting meaningful insi ghts ~Easy integration of data from various sources }

Which type of analytics focuses on predicting future trends?
{ ~Descriptive analytics =Predictive analytics ~Prescriptive analytics ~Diagnostic analytics }

What does the term "Hadoop" refer to in the context of big data?
{ ~A programming language for data analysis ~A database management system =An open-source frame work for processing large datasets ~A visualization tool for data analytics }

Which term refers to the process of discovering patterns and insights in large datasets?
{ =Data mining ~Data warehousing ~Data cleaning ~Data validation }

How many V's of Big Data?
{~2 ~3 ~4 =5}

In Big Data environment, Veracity of data refers to?
{ ~Quality or fidelity of data ~Large size of the data that cannot be process ~Small size of the data that ca n easily process =All of the mentioned above}

Data in which bytes of size is called Big Data?
{ ~Tera ~Giga =Peta ~Meta }

Amongst which of the following shows an example of unstructured data?
{ ~Students roll number, age ~Videos ~Audio files =Both videos and Audio Files }

Amongst which of the following is/are not Big Data Technologies?
{ ~Apache Hadoop ~Apache Spark ~Apache Kafka =Apache Pytarch }

Amongst which of the following can be considered as the main source of ...
{~Twitter ~Facebook ~Webpages =All of the mentioned above }

According to analysts, for what can traditional IT systems provide a foundation when they're integrated wit h big data technologies like Hadoop?
{ =Big data management and data mining ~Data warehousing and business intelligence ~Management of Hadoop clusters ~Collecting and storing unstructured data }

Data in ____ bytes size is called big data
{ ~Tera ~Giga =Peta ~Meta }

What are the main components of big data?
{ ~HDFS ~MapReduce ~YARN =All of the above }

The total forms of big data is ____
{ ~1 ~2 =3 ~4}

Transaction of data of the bank is a type of.
{ =structured data ~unstructured data ~Numerical data ~Semistructured data }

_____ is a collection of data that is used in volume, yet growing exponentially with time
{ ~Database =Big Data ~RDbMS ~Non of the above}

What are the different features of Big Data Analytics?
{~opensource ~scalability ~Data Recovery =All the above}

Unprocessed data or processed data are observations or measurements that can be expressed as text, n umbers, or other types of media.
{=True ~False}

In Big Data environments, Velocity refers –
{~Data can arrive at fast speed ~Enormous datasets can accumulate within very short periods of time ~

Velocity of data translates into the amount of time it takes for the data to be processed =All of the mentio ned above}

Which of the following are Benefits of Big Data Processing?
{ ~Cost Reduction ~Time Reductions ~Smarter Business Decisions =All of the mentioned above }

_____ involves the simultaneous execution of multiple sub-tasks that collectively comprise a larger task.

{ =Parallel data processing ~Single channel processing ~Multi data processing ~None of the mentioned above }

MongoDB is a ____ database.
{ ~SQL ~DBMS =NoSQL ~RDBMS }

Big data analysis does the following except?
{ ~Spreads data =Analyze data ~Organizes data ~Collect data }

Which of the following can be generally used to clean and prepare big data.
{ ~Pandas ~SQL ~NOSQL =Data Warehouse}

What is the use of data cleaning?
{ ~To remove the noisy data ~Transformations to correct wrong data ~correct the inconsistencies in data =All of the above }

Choose the languages which are used in data science.
{ ~c ~c++ =R ~Ruby}

Choose whether the following statement is true or false: Unstructured data is not organized
{ =True ~False ~May be true or false ~cannot be determined }

Machine learning is a subset of which of the following.
{ =Artificial intelligence ~Data learning ~Deep Learning ~None of the above }

Identify the key data science skills among the following
{ ~Data Visualization ~Machine Learning ~Statistics =All of the above }

Choose the correct components of data science.
{ ~Domain Expertise ~Data Engineering ~Advanced Computing =All of the above }

What is the main characteristic that distinguishes "Big Data" from traditional data processing?
{ ~Smaller data volume ~Simpler data structures ~Faster data processing =Large data volume }

Which of the following is NOT one of the commonly recognized "Vs" of Big Data?
{ ~Volume ~Velocity =Veracity ~Variety }

What is the role of "ETL" in a Big Data environment?
{ =Extract, Transform, Load - The process of moving data between databases ~Event Time Logging - Ca pturing timestamps of data events ~Elastic Tensor Layer - Handling complex tensor computations ~Ephe meral Task Lifecycle - Managing short-lived data processing tasks }

Which term refers to the process of discovering meaningful patterns and insights from data?
{ ~Data visualization ~Data preprocessing =Data analytics ~Data scrubbing }

What does the acronym "NoSQL" stand for in the context of databases used in Big Data environments?

{ ~Non-Sequential Query Language ~Non-Supervised Query Logic =Not Only SQL ~New Order of SQL }

Which term describes the process of preparing and cleaning raw data for analysis?
{ ~Data visualization ~Data aggregation =Data transformation ~Data interpretation }

"Predictive Analytics" is most closely related to which question?
{ ~What happened? ~Why did it happen? =What will happen? ~What should we do about it? }

Which type of analytics involves using historical data to forecast future trends and outcomes?
{ ~Prescriptive Analytics ~Diagnostic Analytics =Predictive Analytics ~Descriptive Analytics }