**BIG DATA ANALYSIS WITH IBM CLOUD**

**COMPUTING**

**312621104026: Santhosh S**

**Phase-3 Document submission**

****

**The Objective:**

Big Data Analytics accelerates the process by which

firms transform information into insight. Making educated

decisions about products, operations, marketing, and other

company objectives follows from these insights. Large data

sets need storage, which may be costly to maintain.

**Introduction:**

IBM offers a wide range of cloud-based big data and analytics solutions that can help you gain insights from your data. Here are some resources to get started:

Big data analytics and cloud computing are two of

the most important technologies in the IT industry. Big data

refers to the storage, processing, and analysis of large amounts

of data, while cloud computing is about providing the

infrastructure to enable such processes in a cost-effective and

efficient manner

**Development:**

IBM offers a comprehensive guide for designing and deploying cloud-based data and analytics solutions called "Building Big Data and Analytics Solutions in the Cloud" ¹. It covers the key concepts, architectures, technologies, and best practices of IBM cloud offerings and services. The guide also provides real-world case studies and examples to illustrate how to leverage the power and flexibility of IBM cloud for data and analytics.

In general, big data projects can be broken down into five phases: ad-hoc exploration, use-case discovery, opportunistic development, data management, and optimization ⁴. Here are ten points to consider when developing a big data analytics project with IBM cloud computing:

**1. Identify the business problem you want to solve:**

This is the first step in any big data project. You need to identify the business problem you want to solve with your analytics solution.

* Define your objectives: Identify the specific business problem you want to solve and the goals you want to achieve through big data analysis.
* Assess your data :Determine what data you have available and what data you need to collect to solve your business problem.
* Choose the right tools: Select the appropriate big data and analytics tools that can help you achieve your objectives.
* Develop a plan: Create a plan that outlines how you will use big data analysis to solve your business problem.
* Implement your solution: Execute your plan and implement your big data solution.

**2. Define the scope of your project:**

Once you have identified the business problem, you need to define the scope of your project. This includes identifying the data sources you will use, the analytics tools you will use, and the expected outcomes.

* Identify the business problem: Define the specific business problem you want to solve with big data analysis.
* Assess your data: Determine what data you have available and what data you need to collect to solve your business problem.
* Choose the right tools: Select the appropriate big data and analytics tools that can help you achieve your objectives.
* Develop a plan: Create a plan that outlines how you will use big data analysis to solve your business problem.
* Implement your solution: Execute your plan and implement your big data solution.

**3.Identify the data sources you will use:**

You need to identify all the data sources that are relevant to your project. This includes both internal and external data sources.

* Data availability: Determine what data is available to you and what data you need to collect to solve your business problem.
* Data quality: Assess the quality of the data you have and determine if it is suitable for your analysis.
* Data relevance: Identify the data sources that are most relevant to your business problem.
* Data volume: Determine the amount of data you need to analyze and ensure that your big data solution can handle it.
* Data variety: Consider the different types of data that you need to analyze, such as structured, semi-structured, and unstructured data.

IBM offers a range of big data and analytics solutions that can help you collect, store, and analyze your data.

**4.Determine the data quality requirements:**

You need to determine the quality of the data that you will be using for your project. This includes identifying any missing or incomplete data.

* Data accuracy: Determine the level of accuracy required for your analysis and ensure that your data meets those requirements.
* Data completeness: Ensure that your data is complete and contains all the necessary information required for your analysis.
* Data consistency: Ensure that your data is consistent and free from errors or discrepancies.
* Data timeliness: Determine how frequently your data needs to be updated and ensure that it is updated accordingly.
* Data relevance: Ensure that your data is relevant to your business problem and analysis.

**5.Choose the appropriate analytics tools:**

You need to choose the appropriate analytics tools for your project. This includes selecting tools for data visualization, machine learning, and predictive analytics.

* IBM Analytics Engine: This is a fully-managed big data and analytics service that provides Apache Spark and Apache Hadoop clusters. It enables you to quickly provision and scale clusters, and provides a range of tools for data processing, machine learning, and more.
* IBM Watson Studio: This is a cloud-based platform that provides a suite of tools for building, training, and deploying machine learning models. It includes features such as autoAI, which automates the process of building machine learning models.
* IBM Cloud Pak for Data: This is an all-in-one cloud-based data and AI platform that provides a unified analytics experience. It includes a range of tools for data preparation, building models, and deploying them into production.
* IBM Streams: This is a real-time analytics platform that enables you to analyze large volumes of streaming data in real-time. It includes a range of tools for data processing, machine learning, and more.

**6. Develop a data model:**

You need to develop a data model that will be used for your project. This includes defining the structure of your data and how it will be stored.

**7.Create a data pipeline:**

You need to create a data pipeline that will be used for your project. This includes defining how your data will be ingested, processed, and stored.

* Identify data sources, including databases and external APIs.
* Utilize IBM Cloud services like IBM Cloud Object Storage for secure data storage.
* Implement efficient data ingestion processes for bringing data into the analytics environment.
* Clean, transform, and enrich data using tools like Apache Spark or IBM Watson Studio.
* Analyze data using tools like IBM Watson Studio, and visualize insights using platforms like IBM Cognos Analytics.

**8. Implement your analytics solution:**

You need to implement your analytics solution using the appropriate tools and technologies.

* .Data Collection: Gather relevant data from various sources. This could include structured data from databases and unstructured data from social media, sensors, etc.
* Data Ingestion: Transfer the collected data to the IBM Cloud platform. IBM Cloud provides services like IBM Cloud Object Storage and IBM Cloud SQL Query for managing and querying large datasets.
* Data Processing: Use tools like Apache Hadoop and Apache Spark to process and analyze the data. IBM Cloud offers services like IBM Analytics Engine and IBM Cloud DataFlow for these purposes.
* Data Storage: Store the processed data efficiently. IBM Cloud provides various storage options such as IBM Cloud Object Storage, IBM Db2 on Cloud, and IBM Cloud Databases for different data types and use cases.
* Machine Learning: If your analytics solution involves machine learning, IBM Watson Machine Learning can be used to build, train, and deploy machine learning models.

**9. Test your solution:**

You need to test your solution to ensure that it is working as expected.

* Unit Testing: Test individual components of your solution in isolation. This ensures that each part of your analytics pipeline works as expected.
* Integration Testing: Test how different components work together. Verify that data flows correctly between components and that integration points function as intended.
* Data Quality Testing: Check the quality of your data. Ensure that the data is accurate, complete, and consistent.
* Performance Testing: Evaluate the performance of your solution under different conditions.
* Regression Testing: After making any changes or updates to your solution, perform regression testing to ensure that existing functionality still works as expected.
* Security Testing: Assess the security measures of your solution. Identify and address vulnerabilities related to data storage, data transmission, access controls, and authentication mechanisms. Security testing is vital, especially when dealing with sensitive or personal data.
* User Acceptance Testing (UAT): Involve end-users or stakeholders to validate the solution.

**10. Monitor your solution's performance:**

You need to monitor your solution's performance over time to ensure that it continues to meet your business needs.

* Establish Key Performance Indicators (KPIs): Define specific metrics that align with your solution's goals. These could include processing speed, data accuracy, system uptime, resource utilization, and user satisfaction. KPIs provide a clear benchmark for evaluating your solution's performance.
* Real-Time Monitoring: Implement real-time monitoring tools to track your solution's performance as it happens. This allows you to identify and address issues promptly. Services like IBM Cloud Monitoring and IBM Cloud Log Analysis can help you gain insights into your system's behavior in real time.
* Log Analysis: Regularly analyze logs generated by your solution. Logs can provide valuable information about errors, user activities, system behavior, and security events. Use log analysis tools to identify patterns, anomalies, and potential areas for improvement.
* Alerts and Notifications: Set up automated alerts and notifications based on predefined thresholds. When a metric deviates from the expected range, these alerts can notify your team, enabling them to take immediate action before issues escalate.
* Capacity Planning: Monitor resource usage such as CPU, memory, and storage. Use this data to plan for future capacity needs. Scalability is crucial, especially if your solution experiences varying workloads.
* Security Monitoring: Implement security monitoring tools to detect and respond to security threats. Monitor access logs, authentication attempts, and data transfers to identify suspicious activities. IBM Cloud provides security services like IBM Cloud Security Advisor to help you enhance your solution's security posture.

**Conclusion:**

Implementing big data analysis on IBM Cloud Computing empowers businesses with scalable, secure, and efficient data processing capabilities. Leveraging IBM's advanced tools and services, organizations can derive actionable insights, enhance decision-making, and drive innovation. The seamless integration of data collection, processing, and analysis ensures a streamlined analytics workflow. Real-time monitoring and proactive issue resolution guarantee optimal performance and user satisfaction.