#SFL Scientific

**1. What is a Data Lake? Explain its benefits, how it differs from a data warehouse, and how it might benefit a client.**

**A.** A data lake is a storage repository that holds a vast amount of raw data in its native format until it is needed. While a hierarchical data warehouse stores data in files or folders, a data lake uses a flat architecture to store data. Each data element in a lake is assigned a unique identifier and tagged with a set of extended metadata tags. When a business question arises, the data lake can be queried for relevant data, and that smaller set of data can then be analyzed to help answer the business questions. The term data lake is often associated with Hadoop-oriented object storage. In such a scenario, an organization's data is first loaded into the Hadoop platform, and then business analytics and data mining tools are applied to the data where it resides on Hadoop's cluster nodes of commodity computers.

Benefits of Data Lake -

1. Scalability – It is the capability of a data system, network, or process to handle a growing amount of data or its potential to be enlarged to accommodate that data growth. One of the horizontal scalability tools is Hadoop, which leverages the HDFS storage.

2. The Schema – The traditional data warehouse will not support schema less storage. But the Data Lake leverages the Hadoop simplicity to store the data based on schema less write and schema based read mode, which is very much useful at the time of data consumptions.

3. Converge All Data Sources – Hadoop powered to store the multi structured data from diverse set of sources. In simple words the Data Lake has ability to store logs, XML, multimedia, sensor data, binary, social data, chat, and people data.

4. High Speed Data – In order to have the high-speed data in the Data Lake, it should use few of the tools like Scribe, Kafka, and Flume which can acquire and queue the high-speed data. By leveraging this high-speed data can integrate with the historical data to have its fullest insights.

5. Advanced Analytics: Unlike a data warehouse, the Data Lake excels at utilizing the availability of large quantities of coherent data along with deep learning algorithms to recognize items of interest that will power real-time decision analytics.

**2. Explain serverless architecture. What are its pros and cons?**

**A.** A serverless architecture is a system of constructing and running applications and services without infrastructure management. To cut a long story short, when someone uses the term ‘serverless architecture’ or ‘serverless computing’ if you will, they are describing applications that rely on a mix of third-party services, custom code (that runs in ephemeral containers). These serverless apps are event-driven and are counting on cloud-hosted remote procedure calls — FaaS (Functions as a Service). It’s referred to as “serverless computing” for a very simple reason. Anyone who owns the system is not required to pay for servers for the back-end code to run. The application is, of course, still running on servers, albeit any hosts like Amazon Web Services, Google Cloud Platform, etc. carry out all server management. Therefore, in order to run the application, users circumvent certain procedures such as provision and server maintenance.

Pro’s -

1. Scalability - Serverless architectures are inherently scalable allowing your application to, for example, scale automatically as the user base grows or usage increases.

2. Faster delivery - Function as a Service (FaaS) functions are a lot more straightforward when deploying, as opposed to the deployment of a huge server.

3. Reduced maintenance - With no backend infrastructure to manage, as it all falls to Cloud vendors, you are free of maintenance problems you’d have if you’re running a server on your own. That includes updates, patches and taking care of potential hacks.

Con’s -

1. Limits and cost alterations - They are prone to change based on your previous business scope history with Cloud vendors for example.

2. Security - With many components the total surface open to hazardous attacks is significantly larger, and with reduced overall control you can be blindsided when it happens.

3. Performance - All of the above + few technical bits like increased request latency influence performance and user experience.

**3. Please provide a diagram of the ETL pipeline from Section 1 using serverless AWS services. Describe each component and its function within the pipeline.**

**A.**

Diagram

Description automatically generated

**4. Describe the process of deploying infrastructure as code on AWS or Azure and why this is a good method of architecture delivery for clients.**

**A.** Infrastructure as Code (IaC) is the managing and provisioning of infrastructure through code instead of through manual processes. With IaC, configuration files are created that contain your infrastructure specifications, which makes it easier to edit and distribute configurations. It also ensures that you provision the same environment every time. By codifying and documenting your configuration specifications, IaC aids configuration management and helps you to avoid undocumented, ad-hoc configuration changes. Version control is an important part of IaC, and your configuration files should be under source control just like any other software source code file. Deploying your infrastructure as code also means that you can divide your infrastructure into modular components that can then be combined in different ways through automation. Automating infrastructure provisioning with IaC means that developers don’t need to manually provision and manage servers, operating systems, storage, and other infrastructure components each time they develop or deploy an application. With AWS CloudFormation Platform can be used to describe the desired state of your infrastructure, which the tool can then provision. One can also use this platform for configuration management to maintain your systems in the desired state.

Text

Description automatically generated with medium confidence

Benefits -

1. Reducing shadow IT - Much of the shadow IT within organizations is due to the inability of IT departments to provide satisfactory and timely answers to operational areas, especially around IT infrastructure and systems enhancements. Shadow IT poses significant security risks as well as potential unforeseen costs for the organization. Enabling a fast response to new IT requirements through IaC assisted deployment assures higher security and compliance with corporate IT standards and helps with budgeting and cost allocation.

2. Improving customer satisfaction - Being able to deliver a quality service component within a short period of time contributes to customer satisfaction and improved perception of IT within an organization.

3. Standardizing - Coding the creation of new infrastructure ensures a consistent set of instructions and standardization. Manual configurations are prone to errors and minor changes which can create ever so slight differences that, over time, result in major non-conformities with the standard (and technical debt).

**5. Describe modern MLOps and how organizations should be approaching management from a tool and system perspective; what tools do you recommend? Where would you provide such services in your assessment?**

**A.** With increasing complexity in the context of modeling and frameworks as well as the inherent computational complexity involved with these complicated models, organizations are finding it harder and harder to keep up with the evolving needs of the ML world. This is where MLOps come into play and can offer your organization the structure you need to be successful in executing on your vision and mission. MLOps is the fusion of traditional DevOps processes in the context of data science and machine learning. ML processes are data-centric contrasted with the code-centric philosophy of DevOps. Taking learnings and methodologies from DevOps and applying them to the context of Data and ML yields an operating model termed MLOps.

Benefits of MLOps for an organization -

When it comes to business value, the question organizations typically ask is about time to value: how can I optimize and operationalize this process to achieve more? In terms of MLOps, it means figuring out how to speed up the life cycle of taking a model from lab to production and derive business value to inform insights and intelligent automation recommendations—translating data into actionable decisions and meaningful business outcomes. Once you define and streamline your internal processes, data-driven intelligence can be embedded into the process automation life cycle. That’s where MLOps can shine. Embedding MLOps in a true outcome-driven process will remove the manual process out of your systems to help you move ML models across your applications through testing, training and go-live. This will support the continuous improvement life cycle to combat model degradation. As your organization matures its machine learning, process and automation workflows will reap the benefits from concept to realization of ML deployments. I would recommend to AWS SageMaker, Google Cloud AI or Data Robot to the organization to deploy machine learning models at any scale.