SFL Scientific Solution

1. **What is a Data Lake?**
   1. **Data Lake:**
      1. A data lake is a central location that holds a large amount of data in its native or raw format.
      2. It often consolidates all an organization’s data in one single central location, where it can be saved “as is” without the need to impose a schema.
      3. It stores data from all stages of the refinement process starting from the staging layer to compute layers.
      4. And it can store all types of data – including unstructured and semi-structured data like image, video, audio, etc.
      5. Example – Azure gen lake 2.
   2. **Benefits of Data Lake and how it’s different from Datawarehouse.**
      1. Datawarehouse can process only structured data whereas data Lake can store structured, semi-structured and unstructured data.  
         Benefit for the client: They need not maintain different types of databases for structured and unstructured data.
      2. Datawarehouse can scale up to a certain extent and becomes exponentially more expensive, but data Lake can scale to any amount with less cost.  
         Benefit for the client: Scalability with low cost
      3. Data warehouse is mostly meant for the business professional but data Lake can be used by business professionals and data scientists.  
         Benefit for the client: A unified space for everyone.
2. **What is serverless architecture? And what are the pros and cons?**
   1. Serverless is a cloud computing execution model where the cloud provider dynamically manages the allocation and provisions of the server.  
      A serverless application runs in stateless compute containers that are event-triggered and are fully managed by event triggers.
   2. The pricing is based on the number of executions rather than pre-purchased compute capacity.
   3. Example: Azure functions, AWS Lambda.
   4. **Pros:** 
      1. Cheaper than traditional cloud
      2. It is scalable to any capacity.
      3. It lowers human resource costs.
      4. It focuses more on user experience.
   5. **Cons:**
      1. Third-party has control of the execution.
      2. It has learning curve.
3. **ETL Flow:**  
   Diagram, table

   Description automatically generated
4. **Modern ML Ops.**
   1. ML Ops is the fusion of the traditional DevOps process in the context of data science and machine learning.
   2. Graphical user interface, application

      Description automatically generated
   3. **Steps Involved in Data:**
      1. Identify and collect:
         * The First step in a dataOps enabled pipeline is identifying sources and source system where the target data lives.
      2. Process:
         * Once the source data is identified, it needs to be processed and transformed into a form that can be analyzed or explored.
      3. Store:
         * Choosing the technology that meets your organization’s access rights, security, and privacy needs. Using a Data warehouse or a data Lake is a typical option.
   4. **Steps involved in ML:**
      1. Extract:
         * The goal is to extract the feature that contains information about the data which can resolve the business problem.
      2. Train:
         * Training is fitting a model to your train dataset while maintaining optimal performance and meeting the requirement.
      3. Optimize:
         * Use the appropriate method to optimize the model before going live.
   5. **Steps involved in Continuous integration:**
      1. Plan, package and develop:

* When the model is ready to be published, the underlying code, dependency will be pushed to a repo and merged to the main branch.
* The push will trigger the build for the model.
  + 1. Test:
       - After the build is generated and packaging is done and the container is ready to be shipped, the process continues with performing the unit test and integration testing.
  1. **Steps involved in Continuous integration:**
     1. Staging:
        + Staging is a place where the artifact of CD will reside.
     2. Release and configuration:
        + In this step the mode will be deployed to the target system.
     3. Validation and Monitoring.
        + ML model will be validated in this step and monitored.