Data Lake Architecture -

A Comprehensive Design Document

Medical Data Processing Company

# Tracker

## Revision, Sign off Sheet and Key Contacts

## Change Record

|  |  |  |  |
| --- | --- | --- | --- |
| Date | Author | Version | Change Reference |
| 06/04/2020 | FirstName LastName | 0.1 | Initial draft |

## Reviewers / Approval

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Version Approved | Position | Date |
| FirstName LastName | 1.0 | Udacity Reviewer  Enterprise Data Lake Architect |  |

## Key Contacts

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Role | Team | email |
| FirstName LastName | Data Architect | Medical Data Processing | student@email.com |

# Note from Instructor:

# Consider this as a comprehensive design document that you will deliver to the technical audience of the company.

# Provide detailed design and implementation level details

# You are expected to provide at least 6 pages worth of content (Does not include the cover (title) page and tracker page)

# Each section has a set of guiding questions that will help you derive the responses.

# Purpose < approx. ¼ page>

<What is the purpose of the document (Summary)?>

<What does the document contain?>

<Why are you creating this document?>

<Who is the target audience?>

<Inscope and/or out of scope items?>

# Requirements <approx. 1 page>

<Summary of requirements for Data Lake. Summarize your understanding of the problem statement. >

<Existing Technical Environment>

<Current Data Volume>

<Business Requirements>

<Technical Requirements>

<Where do you find these requirements? Have you seen them somewhere before? >

# Data Lake Architecture design principles <approx. ½ page>

<List of the design principles. What is the baseline criteria to design the system? What rules/guiding principles should be followed?>

<Provide rationale of “why” the design principal was selected and how would it help in overall Data Lake design long-term for Medical Data Systems>

[You may not use this example in your final solution]

e.g.: Leverage open source tools as much as possible

# Assumptions <approx. ⅓ page>

<What are the assumptions you have made while creating the Data Lake architecture?>

<Be creative, what questions did you have while designing the architecture?>

<What data is missing in the problem statement, and you made assumptions about it to create the architecture?>

<Describes any potential risks that may be created now or in future based on these assumptions>

[You may not use this example in your final solution] e.g.:

1. Hadoop cluster will use Linux operating system
2. Data Lake will not support X, Y, Z

# Data Lake Architecture for Medical Data Processing Company

< Embed your Architecture Diagram of Data Lake you created in Step 2 >

# Design Considerations and Rationale <at least 3 pages>

## Ingestion Layer

<How do you plan to ingest different types of data?>

<How would you ingest data coming from Databases, FTP servers, APIs?>

<What tools would be used? Why? >

<How would the ingestion layer design scale?>

<What other tools were considered? (3rd party tools, open source tools considered but did not make it to the architecture you are proposing). Are there other shortcomings to your selection of tools? If so what? Does the 3rd party tool solve that?>

## Storage Layer

<How do you plan to store a vast amount of data? >

<How would the system handle 20% YoY Data Growth rate?>

<How do you plan to handle back-up and recovery? What are the strategies?>

<How do you plan to store custom **metadata** information? What type of information would metadata hold?>

<What format of the data do you plan to use? Why?>

<How do you plan to secure data (at a high-level)? Identify 2-3 techniques/tools/considerations>

<What other tools were considered? (3rd party tools, open source tools considered but did not make it to the architecture you are proposing). Are there other shortcomings to your selection of tools? If so what? Does the 3rd party tool solve that?>

## Processing Layer

<How do you plan to process the data?>

<How do you satisfy different processing needs? Batch, Realtime, CDC?>

<How do you enable ad-hoc querying capabilities?>  
<What different tools are involved for processing?>

<What other tools were considered? (3rd party tools, open source tools considered but did not make it to the architecture you are proposing). Are there other shortcomings to your selection of tools? If so what? Does the 3rd party tool solve that?>

<How does the proposed architecture scale with respect to processing?>

## Serving Layer

<What do you mean by serving layer?>

<What type of data do you plan to store here?>

<How would the data in the serving layer be used?>

# 8. Conclusion <approx 2-5 lines>

<Conclude the contents of the document. Provide recommendations on next steps if any.>

# 9. References <If any>

<Provide links of any external documentation, wiki, blogs that you used to complete your research to put this solution together>