TecFusion Consulting

DATA LAKE SOLUTION

Organizations is leading European Telecom Client having Multi-Functional data type kept in 30+ system serving entire business operation.

Data from 30 plus sources been pulled in DW system for further analysis using data integrator solution and enterprise reporting been in place to send various trending reports based on weekly ,monthly quarterly basis. This process has been in place over 4 years as data is growing more and total archival policy of system is 7 years as per GDPR now a demand been raise to load data using faster technique than traditional data integrator solution.

Also more heterogeneous system been added across enterprise and there is a demand of analyzing data from various sources and derive a trend on various aspect of data right from sales analysis to inventory, marketing data, voice calls, customer, supplier data and latest addition is security and networking data log analysis.

Unlike traditional data integrator solution which is capable of keeping structured data in normalized form of database data lake solution can keep structured, semi-structured, unstructured and real time data in a single distributed filesystem from various source

Business Requirement & Challenges

- This solution will deliver data-lake solution extracting data from 30 plus source system and putting in HDFS Hadoop system for consolidation and further analysis for trending to satisfy multi-dimensional aspect of corporate need.
- HDFS file system been cost optimised storage solution as opposed to tradition SAN solution with the feature of data high availability, easy connectivity of data transformation tool /micro services to translate data into a meaningful analytical dataset which can then be used to derive story about the data using various graphical tool like Power BI,Tableau etc.
- As opposed to traditional SAN S3 storage is cost effective and support any format of data

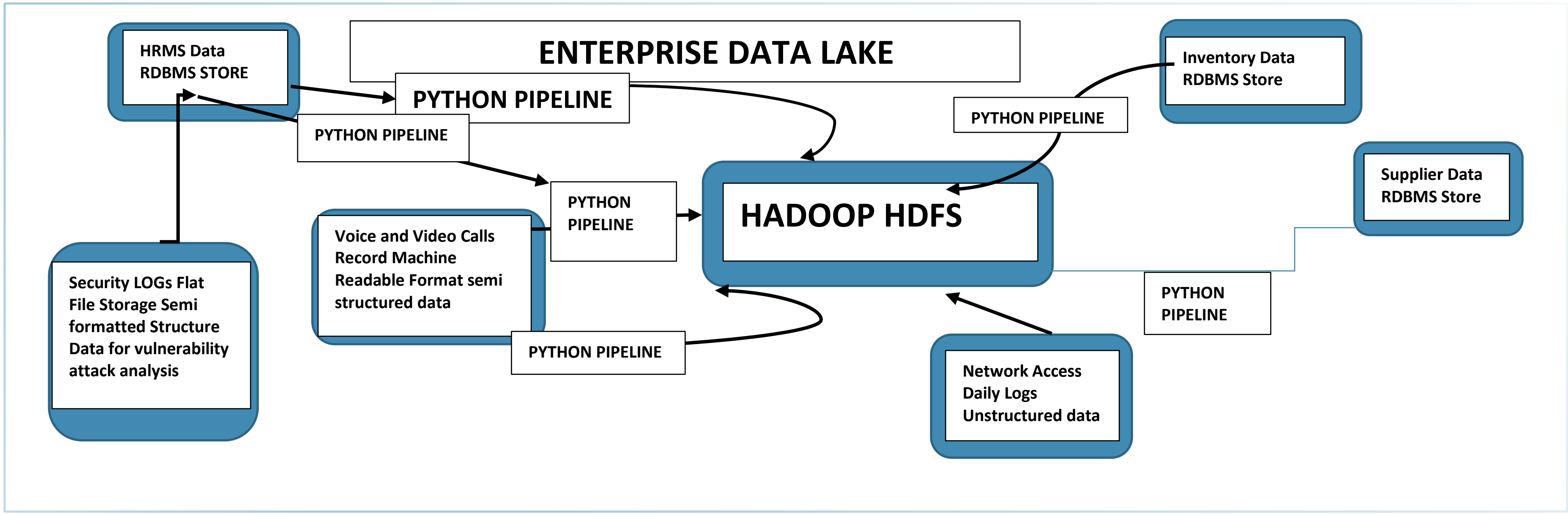
Approach:

The objectives of this project are to

- Provide a centralised platform for creating data-lake from various structured, semi-structured and unstructured sources.
- Creating a data pipeline using python code to ext   ract data from various source system.
- Data from source system are in different format like CSV, unstructured security logs, RDBMS.
- Hadoop HDFS system has been chosen to put structure, semi structure data from 30 plus source
- Python at source been used for cleansing data and reformatting.
- Python data pipeline has been chosen to build up data extraction from various sources which can be extended in future to extract data from real time source system.

Solution:

Figure below illustrates the Data Lake build from heterogeneous source system.


HRMS Data
RDBMS STORE
ENTERPRISE DATA LAKE
PYTHON PIPELINE
Inventory Data
RDBMS Store
PYTHON PIPELINE
PYTHON PIPELINE
Supplier Data
RDBMS Store
PYTHON
PIPELINE
HADOOP HDFS
Voice and Video Calls
Record Machine
Readable Format semi
structured data
Security LOGs Flat
File Storage Semi
formatted Structure
Data for vulnerability
attack analysis
PYTHON
PIPELINE
PYTHON PIPELINE
Network Access
Daily Logs
Unstructured data