***Moved to Cloud What next?***

**In the modern era of distributed computing –** *picking the right tool set is essential*, in this article lets discuss the solution approach for one of customers who is using Snowflake on Azure Cloud

**Current Architecture:**

A Typical Traditional Datawarehouse Architecture - which is widely adapted and practiced across organizations.

Data from Heterogeneous sources [Flat files, CRM, ERP and transaction Databases like SQL-server] sourced from both on-premises data center as well as from cloud provider (Azure Cloud) are extracted, transformed and loaded into Snowflake Data Cloud for Data Analytics Graphical user interface, diagram

Description automatically generated

**Tool Stack used:**

* Talend (ETL)
* Snowflake Hosted on Azure (Datawarehouse)
* Power-BI(Reporting)

**Pain points:**

* Data Onboarding/Latency
* Data Quality
* Metadata Management
* Data Classification and protection
* Performance challenges

**Data Onboarding/Latency**

Customer’s Data is spread across multiple sources residing on both on-Premises and cloud infrastructure and retrieving the datasets (Flat files, CRM, ERP and Transaction Databases) and keeping up with the network latencies is a daunting task in the world of speed and agile environment

**Data Quality**

Due to lack of upfront data quality framework and mechanisms, data anomalies are discovered at the last mile (BI-Reporting) which brings up another challenge about the data correctness and rift between the Data Engineers and Analysts

**Metadata Management**

Working with multiple teams within an organization is always a challenge by itself and meaning of data definitions changes between the teams unless we have a pre-defined metadata and data lineage guidelines

**Sensitive Data and Protection**

“Prevention is better than Cure “- In the era of Modern Cloud, unless we pay special attention to Data sensitivity and protection, there is always a chance to expose the sensitive information like PII and PCI within the organization as well as outside of the organization, which warrants hefty fines and numerous Audits unless we classify the data sets and take appropriate measures to protect the sensitive data

**Performance Challenges**

Often, defining frameworks and choosing the right tools will not alone suffice to meet the Business SLA’s. After all tools will also have certain limitations, we as Data Engineers, Analysts and Architects should understand the nitty-gritty details of the modern Data tool stack and Fine-Tune the complex queries during the migration process as much as possible rather than a mere Lift and Shift approach

**Proposed Solution Architecture -**

In today’s dynamic data world, Data space is crowed with plethora of tools sprouting up like tiny mushrooms in no time and it’s really an overwhelming task for Architecture teams to keep up with the choice of tools available on Open-Source as well as proprietary.

**Guiding Principles:**

* ***Robust Design*** *– Gather the all the requirements, problem statements and edge cases to be solved*
* ***Configurable framework*** *– Build once and adapt to new functionalities with out major code changes*
* ***Future proof transparent solution****- farsighted vision to solve not just current problems as well as unforeseen issues along the way*

**Proposed Solution Approaches:**

Always propose multiple flexible solutions with Customers and guide them to pick the best-fit Solution

**Re-Engineering**

* + Re-Engineering is essential to leverage the Cloud Platform to its best use
  + One-Time investment and long-term benefits and Big Bang approach
  + Time, skillsets, technology stack and Cost are the key factors to consider for Re-Engineering
  + Adapt Out of Box thinking and Multi-Cloud Approach to sustain longevity and less dependent on single Vendor/Cloud Provider.
  + Framework/Configuration based design solution to address the agility, flexibility, and Time to Market needs.

**Refactoring:**

* + Addresses the current problems of the architecture
  + Short term gains and less time consuming in comparison to Re-Engineering
  + May address some of the technical debt but limited to address the future scalability and other challenges
  + Balanced approach -Liberty to pick and choose to “what stays AS IS "and “What needs to be Refactored “

**Architecture Re-Engineering solution**

Diagram

Description automatically generated

**Tool Stack Proposed**

* ~~Talend~~ Python + Snowflake Inbuilt functionalities [Snowpipe, COPYINTO,Streams and Tasks](ETL)
* Snowflake Hosted on Azure (Datawarehouse)
* Power-BI(Reporting)

**How the Pain points Addressed?**

* **Data Onboarding/Latency** 
  + Moving the Data files into Cloud object Storage is cost effective
  + Migrating transactional Database into Cloud solves the scaling and network latency issues to a greater extent
  + Bonus point – If you choose to select snowflake storage and Cloud provider in this case its Azure Cloud, it’s a match made in heaven, eliminates latency, and brings in flexibility, speed and much more
* **Data Quality**
  + Data quality framework is always a team sport, it’s an amalgamation of business and Data teams to define, derive, divide, and conquer the data sets to install the guardrails of quality on data
  + Always remember – *Garbage In is Garbage Out*, unless we distill and refine the data in the raw layer, it’s always a challenge and potential blocker for the last mile course correction
* **Metadata Management**
  + *Do things the Right way from the get-go* which means figure out the data definitions during the data modelling itself and attached the metadata
  + Define the object tagging (Snowflake feature) so that it will be a breeze to find out the field names and what is the true meaning of it
  + Opt for a configurable framework to traverse through and find the data lineage for the objects and fields underneath
* **Data Classification and Protection**
  + Data Classification is pivotal to any organization, and we have tools available in the market to do the job
  + In our case we leveraged the in-built snowflake Data Classification Procedures to classify the sensitive data sets and applied Dynamic Data masking (again in-built Snowflake feature) to protect the data from unauthorized Roles and Users
* **Performance challenges**
  + Identify the bottlenecks prior to the migration, this avoids the future technical debt
  + Snowflake is a SaaS application with Pay per Use Model, we got be cost conscious on storage and compute monitoring
  + Some key considerations for optimization
    - * Setting up Account level parameters
      * Setting up Resource Monitoring and threshold limits on consumption
      * Choosing the Right warehouse type and size for the job
      * Choice of Table types for optimal storage etc

**Closing thoughts**

*“There is always room for improvement with any ideal/ class of the art Solution Architecture “*

*Obtain constant feedback from customers with openness and deliver with confidence and consistency. Modern Data Stack solutions been and shall always be fluid, Tools and methodologies will come and go but some may stay longer than others. Just do not get carried away by the slick design and fancy GUI Evaluate all possible scenarios and that special edge cases!!!*

*“Devil is always in the Detail”*

*Happy learning*

*Raj Gajula*