International Conference on Artificial Intelligence and Machine Learning
In association with International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com) | Volume 8, Issue 5, July-August-2021

Orchestrating Dynamic Big Data End to End ETL Pipeline

Syed Azimuddin Inamdar*[1], Sayyid Abrar[1], Gayatri Bajantri[1]
[1]Department of Computer Science Engineering, VTU, SECAB Institute of Engineering and Technology, Vijayapura, Karnataka, India

ABSTRACT
Now a days data is said to be the new currency and key to triumph. Gathering a rich quality information from numerous dispersed sources across the world necessitates abundant struggle and time. There stand quite a lot of other challenges that consists while transferring information from its start point to its end point. Data ETL pipelines are employed to extend the complete effectiveness of flow of data from its source to the final destination. In the meantime it is automated and decreases the involvement of humans. In spite of prevailing study on ETL pipelines, the study on this topic is limited. ETL pipelines are intellectual representations of end to end data pipelines. To make use of the full possible of the data pipeline, we need to recognize the events that are going in it and the way they're associated in an end to end pipeline. This thesis gives an summary of designing a conceptual model of data pipeline which may be further used as means of communication among various data teams.

Keywords : Bigdata, ETL pipeline.

I. INTRODUCTION

The impartial of this presented thesis is to implement the ETL data pipeline in cloud by optimizing cost and improving the overall performance of the data pipeline. It is a cloud based data integration service that allows us to create data driven workflows for orchestrating data movement and converting data at scale. Using Azure Data Factory we can create and schedule data driven workflows called pipelines that can ingest data from disparate data stores.

ETL stands for Extract, Transform and Load, it is a process of extracting data from one or multiple data sources distributed across the world, then, transforming the data as our business requirements and loading the data into data warehouse. Now a days data is charming more prevalent within the commerce world do to the importance of knowledge crops like APIs, consoles, standards and reports. The data plays important role in policymaking and in the expansion of ML and DL models. Hence, all processes related to data starting from generation of data to data response needs to be watched. The fault detection, reporting and justifying the effect of faults are very multifaceted but unavoidable while construction of effectual data products.

II. EXISTING SYSTEM

Here I have taken two separate cases which existing data pipelines are used, the cases are from telecommunication firm, Each case is separate from other and there is no other interaction between them.

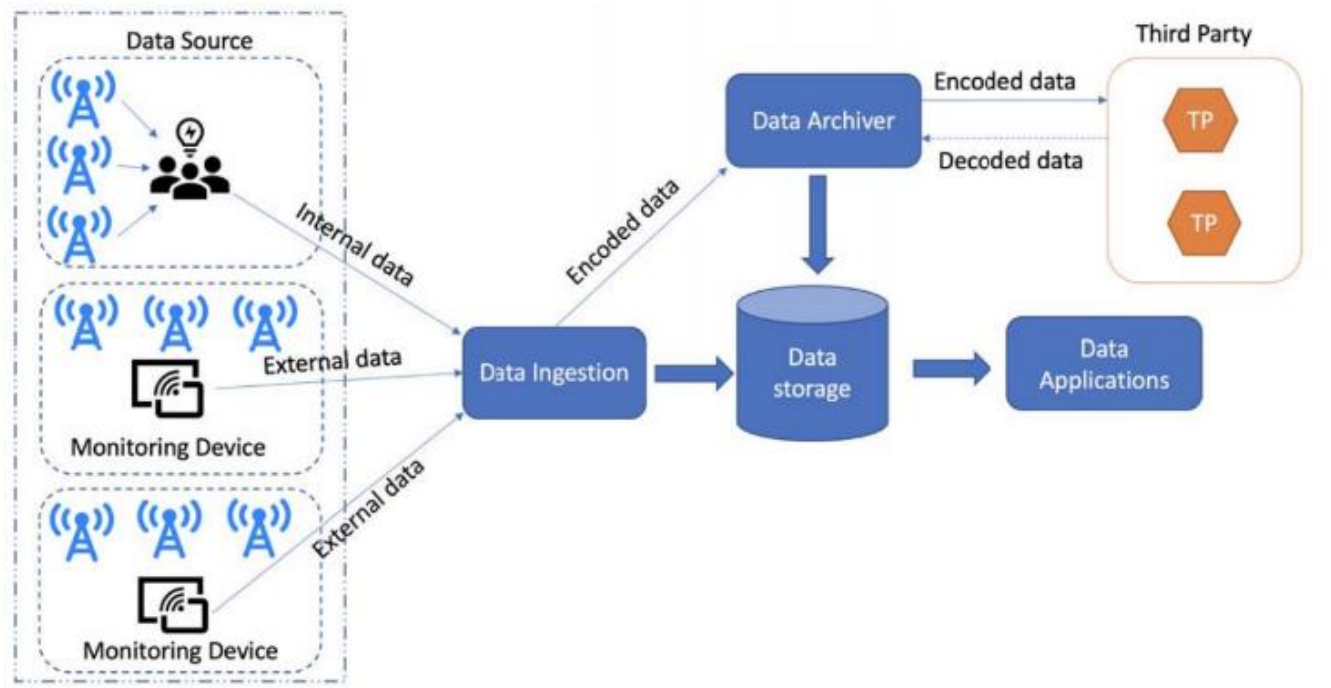

Fig1: Existing pipeline

A. The Data collection process.

The business gathers data from wide spread manifold sources dispersed across the world which are stimulating continuously. Data is composed from the device which is situated in another nation or from the client network. Also the subtle info in the data should be held correctly, it should not leak, Furthermore, the data gathering should deliberate that data generated from different data resources are in different incidences and arrangements. The data can be composed continuously or as batches. The data assortment mechanism should be proficient to adjust with numerous strengths of data-flow. These trials should be addressed religiously when data collection is automated. Fig. 1.5 shows the automatic data collection from the devices. In this case the device is placed inside a piece of equipment owned by customers, the device data is mined without the customer's sensitive information. Base stations have got nodes as well as a device for tracking and managing the nodes. Data assembly equipment are situated on the customer's premises that is physical place which can relate with the nodes straight or with the device to collect the data.

B. Data Governance Pipeline

The data pipeline shown in below fig. 1 is established to aid the data squads in the corporation who are employed with data whensoever it is needed. It is nothing but the connection from where the original data can be copied. There are two types of dumps that data pipeline gets, internal and external. The internal data dump

is the data that is consumed by the squads employed on it inside the corporation and external data dump is the data gathered from the devices in the fields directly. The technique of data absorption is different for different sources and the consumed data is kept in the data storage for future use. The data might have been encrypted connections that need to be decrypted before storing it. There will be a vendor party service to decrypt the encrypted data dump, the data archiver unit sends it to third-party services for decryption. The decoded links are then stored in central storage. Thus the data is made available at any point of the process.

Teams which are working on the data pipeline can appeal data from any phase of the pipeline. The pipeline is tracked manually by flow custodian who is accountable for recognizing the faults in the pipeline and solving them.

C. Problem Statement

Taking into the considerations of my cross case analysis of my data management, I have identified some of the main encounters with the data administration and current data pipeline used in the companies. These are listed below.

Availability of Data: For the successful development of a information product the availability of the accurate data at the accurate time and in the accurate format is very significant. Collection of data is a very tough task and sometimes it fails due to verification failure, climatic factors or even the failure of devices used for collection of data. Even after the gathering of vast data sets from the equipment, it may not reach the defined end point. Data composed can be left unfinished. i.e. unfinished data will be accessible in the data warehouse, it can be due to software letdowns, slices of the data can be vanished. It is hard to identify the loss unless we have a tracking mechanism. Availability of distinct data is vital as input to the model, to scale up the performance.

Quality of Data for systems like ML and DL: Quality of data is vital for systems like Machine Learning and Deep learning. When poor-quality data is nursed to the algorithms of ML and DL, the systems will yield poor-quality output. There should be a clear difference amid faults due to ecological factors and liabilities due to device letdown while collecting the fault logs from the field devices. The key test is when the information is processed to suitable a predefined structure, redundant parts of the data is detached from it. Consequently, we should have a technique to save the original data file. It is not possible to transform the data on the fly and store it, hence it is constantly decent to have the raw data file stored so that it can be retrieved whenever needed by the data squads or when the processed data becomes inadequate to fill the necessities.

Low Storage Capacity: Every team when building their own data pipeline, stores the same data in various forms in the data storage leading to lack of storage space. An amplified quantity of data pipelines will lead to lack of space. When the available storage is separated between different squads, each one will get only a minor share of the real obtainable storage space.

High Cost: To set up a data pipe line in a traditional way, requires high infrastructure cost. Traditionally the pipeline is designed in an on premises system. The system requires high performance processor, RAM and other hardware which make the cost to rise exponentially.

Location Based: Suppose the server is installed in Asia the teams accessing in US will face latency due to geo location of the server. The server responds much faster to the teams accessing the pipeline near to the geo location of the server.

These were the few challenges in data organization and Current data pipelines. Further my thesis I have designed the data pipelines which further reduces or overcomes some of the challenges listed above.

III. PROPOSED SYSTEM DESIGN

The below figure shows the conceptual model, it is a set of ideas used to create data pipelines. The basic fundamentals used to build the data pipeline here are Nodes and Connectors. Connectors are used to interlock the Nodes.

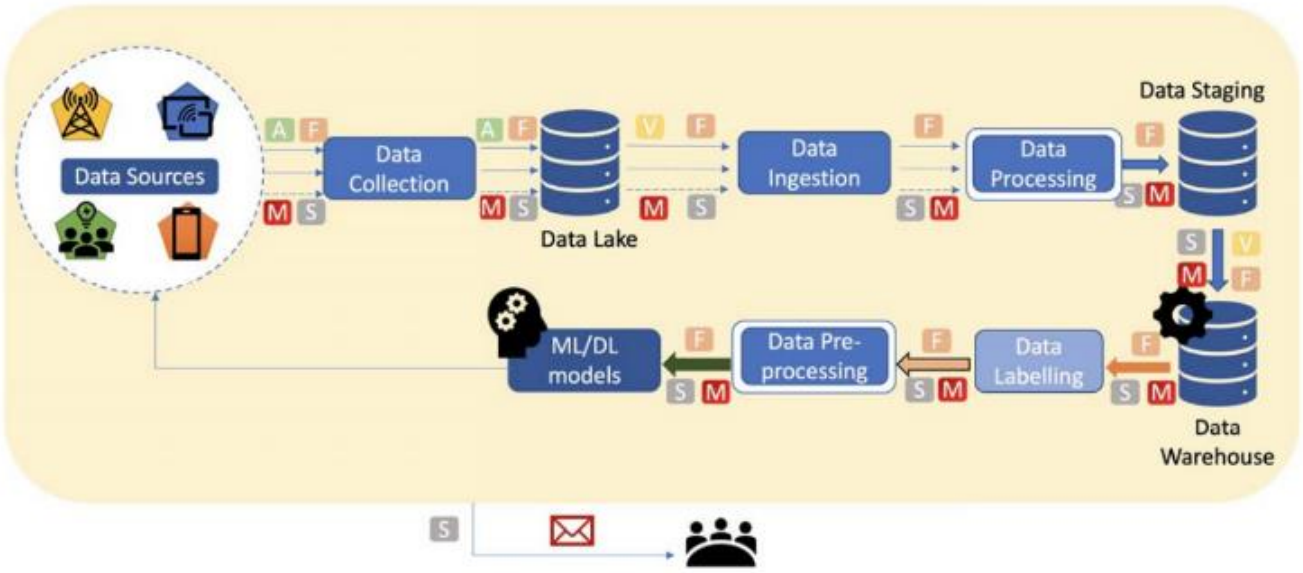

Fig: 2

Data Collection: The information generated from the source could be in the form of batch, intermittent or nonstop type. The three dissimilar flairs of arrows opening from data sources indicate that the connector can carry all the three types of data flow between the data source and data absorption. The data gathering node can gather information from the sources, it can display the consent to gather data from the data sources.

Data Lake: The composed information from the data source will be fresh and it should be kept so that the original files can be repossessed in the future for future use. The data gathering node has to display its right to consume data into the data pipeline. This verification will be approved by connectors among data collection and Data Lake.

Data Processing: Data processing is a complex process which comprises of multiple stages such as data combination, data analysing, data transformation, etc. Data combination is a process in which raw information is translated in a suitable form for arithmetical analysis. The data transformation, is a procedure in which the unstructured combined data is changed into a structured format or semi-structured format. Therefore, we can say that the data processing step translates all diverse kinds of information into a sole format and is stored in data staging area. This is represented in the fig. 1.9 with three dissimilar arrows at the input to data processing showing batch, intermittent and continuous data. The output from the data processing stage is a sole thick arrow as shown.

Data Warehouse: The data staging is a momentary storing area in which the information is stored for authentication. Once the authentication of structured or semi structured information is done, the authenticated information is then located to the data warehouse. This data warehouse acts as a point of admittance from where the information can be reserved for several data applications like formation of report, Machine Learning or Deep Learning applications, etc.

Data Labelling: The data pipeline shows the essential phases to mechanize the data pipeline for Machine Learning applications. Machine Learning algorithms can be of administered, unverified or strengthening. The data labelling phase is done to the administered algorithms, whereas data labelling stage is bounced for unsupervised algorithm. As the most of the businesses are using a administered algorithm method for their Machine Learning applications, the emphasis is high on data labelling.

IV. DISCUSSION AND COMPARISON OF RESULTS

The implementation stage of the mission is where the thorough design is essentially converted into working model. Goal of the stage is to decode the project into a finest conceivable solution in a appropriate programming language. This section covers the application of features of the project, giving particulars of the technology and development atmosphere used. It also gives an impression of the core components of the mission with their phase by phase flow.

The implementation phase requires the following tasks.

1. Cautious planning.
2. Examination of system and limitations.
3. Design of approaches to attain the changeover.
4. Assessment of the changeover technique.
5. Right decisions concerning assortment of the platform.
6. Suitable assortment of the language for application development

The result stage of the project is where the system is assessed and verified in terms of performance and whether or not the goals set in the commencement of the project are attained. Goal of the stage is to attain appropriate data that can be strategized and checked for performance authentication. This section covers the outcomes aspect of the project, giving details of the various complexity stages of the project and relating them.

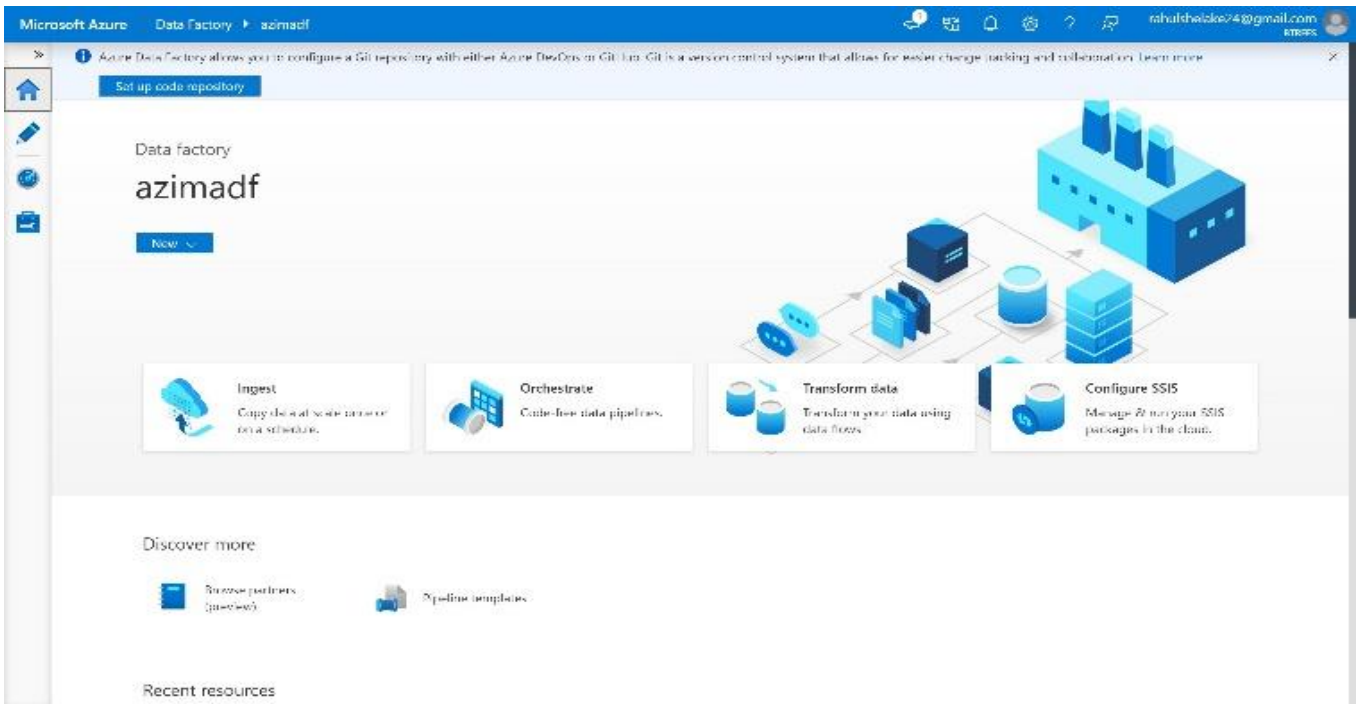

Fig:3 Azure Data Factory

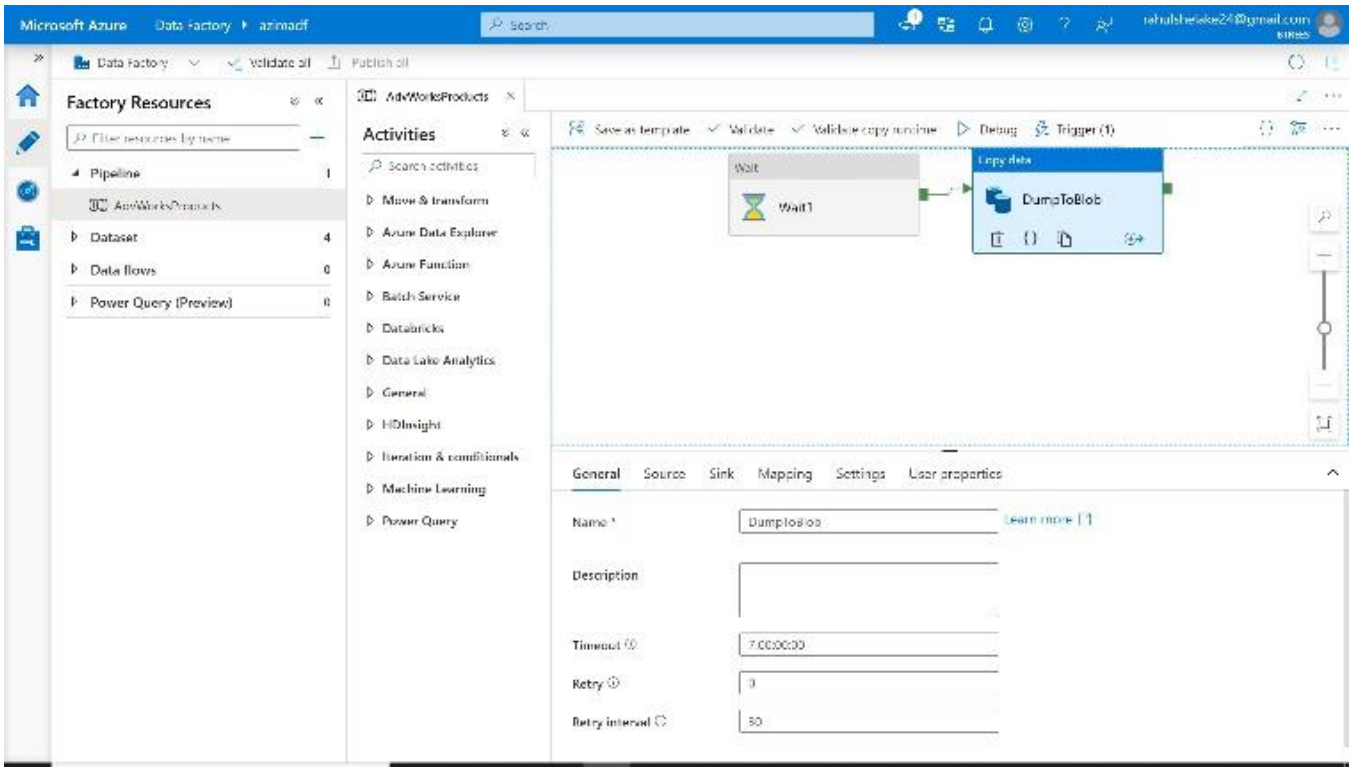

Fig 4: Pipeline

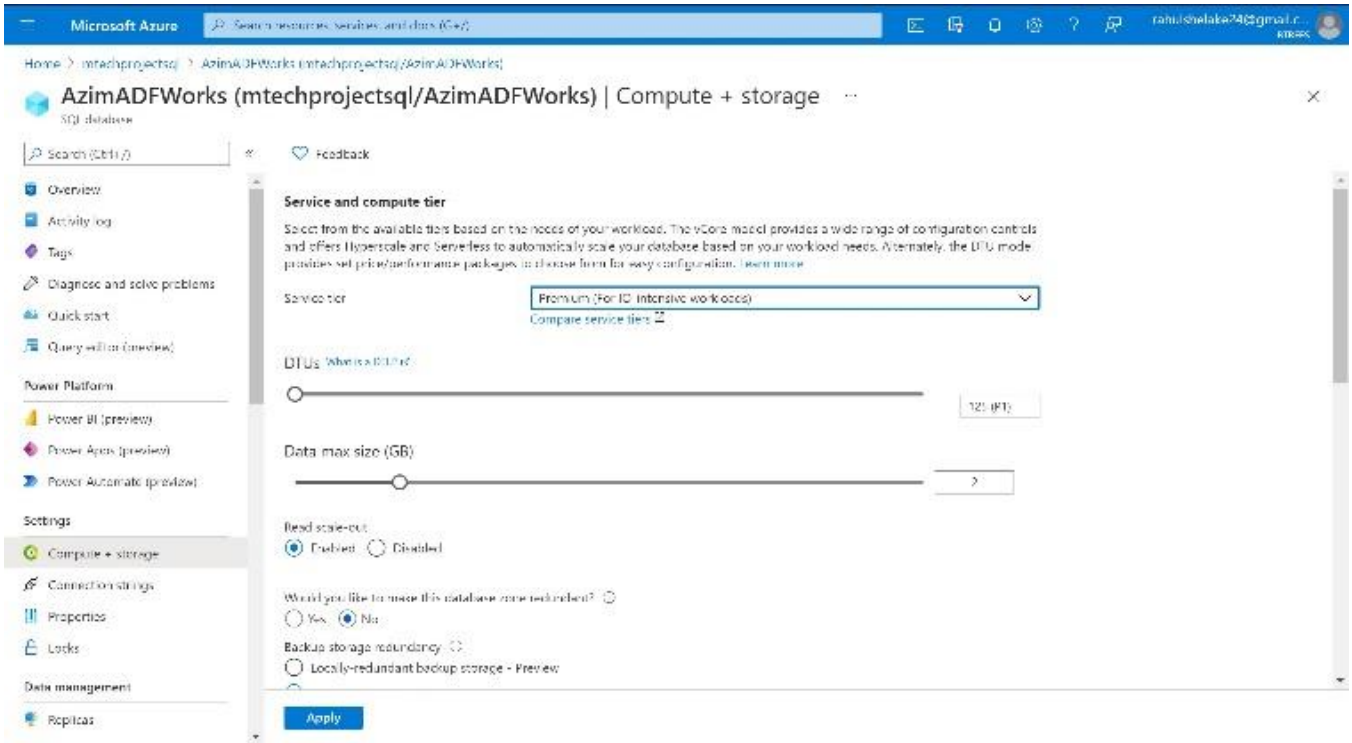

Fig 7:Scaling

Monitoring in Azure Data Factory or any other cloud service like Amazon Web Services (AWS) or Google Cloud Platform(GCP). Orchestration to external services with less or cost optimised way.

Data is new oil, it can manipulate quickly as our need or requirement to achieve much better result in less time using cloud or my design implementation. Using my design, we can extract new value from it as per our future need.

V. CONCLUSION AND FUTURE WORK

It guarantees maximum availability for both single region and multi region databases. It also provides a maximum read availability SLA on multi region databases. To make the automatic failover process more efficient, set a preferred region list for each region. User can access ETL pipeline across the globe without being limited to specific location.

Two other services that can run Spark jobs are Azure Databricks and HDInsight, which can be implemented with more efficient and cost optimized with built-in security provided by cloud services.

VI. REFERENCES

[1]. K. Goodhope, J. Koshy, J. Kreps, N. Narkhede, R. Park, J. Rao, and V. Y. Ye, "Building linkedin's real-time activity data pipeline." IEEE Data Eng. Bull., vol. 35, no. 2, pp. 33–45, 2012.

[2]. E. Deelman and A. Chervenak, "Data management challenges of dataintensive scientific workflows," in 2008 Eighth IEEE International Symposium on Cluster Computing and the Grid (CCGRID). IEEE, 2008, pp. 687–692.

[3]. P. Vassiliadis, "A survey of extract–transform–load technology," International Journal of Data Warehousing and Mining (IJDWM), vol. 5, no. 3, pp. 1–27, 2009.

[4]. J. Trujillo and S. Lujan-Mora, "A uml based approach for modeling ´ etl processes in data warehouses," in International Conference on Conceptual Modeling. Springer, 2003, pp. 307– 320.

[5]. Alkis Simitsis, Kevin Wilkinson, Umeshwar Dayal, Malu Castellanos HP Labs Palo Alto, CA, USA, Optimizing ETL Workflows for Fault-Tolerance, Conference: Proceedings of the 26th International Conference on Data Engineering, ICDE 2010, March 1-6, 2010, Long Beach, California, USA.