IFC-Bank of Italy Workshop on "Data Science in Central Banking: Applications and tools"

14-17 February 2022

Swimming in the data lake: an application to NielsenIQ Homescan[1]

Minnie H Cui, Gene (Fa Gui) Jiang and Botlhale Mosweu,
Bank of Canada

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Swimming in the Data Lake

An application to NielsenIQ HomeScan

Minnie H. Cui[Ψ*], Gene (Fa Gui) Jiang[*] & Botlhale Mosweu[*]

Abstract

We demonstrate the functionality of Azure Data Lakes and cloud computing tools for large data set processing through the case of NielsenIQ HomeScan, a newly onboarded data set at the Bank of Canada. With already more than 61 million observations total from over 40 thousand household panelists, we selected a Data Lake storage solution with a final destination in a Microsoft Azure SQL Database to maintain ongoing quarterly deliveries and to access Azure-based cloud-computing resources. We illustrate the capabilities of Data Lakes for securely and cost-efficiently storing and backing-up data from multiple sources of various formats, structured or unstructured. This flexibility of data formats allows users to apply transformations to data only when needed. Users have the choice of Azure Databricks or various ported software, including JupyterLab, Stata, and R, for analytics. Finally, we show that Databricks significantly reduces runtimes by running on clusters which parallelizes certain processes automatically and auto-scales worker nodes as needed. In our experimentation, Databricks ran 40 times faster than Stata on desktop.

Keywords: cloud computing, Data Lake, Databricks, data science, large data sets

JEL classification: C55, C88

[Ψ] Corresponding author: minniecui@bankofcanada.ca.

[*] Bank of Canada, 234, Wellington Ave., Ottawa ON, K1A 0G9, Canada. The views presented are the authors' and do not represent the official views of the Bank of Canada.

Contents

1. Introduction

As the developed world becomes increasingly connected to digital goods and services, the role of technology in the economy becomes increasingly more important. Observational data sets are now available at much higher frequency and larger scale than traditional surveys. This availability has created opportunities for economists and policymakers to examine economic systems with greater accuracy. However, new methods and technologies are required to take full advantage of these data sets *efficiently*.

Cloud-based technology can significantly improve computational capabilities while bringing cost-efficient storage solutions to large organizations. Azure Data Lakes, for example, can securely store and back-up data of multiple formats, structured or unstructured. More importantly, analytical tools like Databricks are integrated with Azure Data Lakes, reducing data movement while delivering significant performance gains.

In this paper, we discuss our implementation of an Azure Data Lake for the storage and analysis of Canadian NielsenIQ HomeScan data. We discuss an automated data pipeline that handles ingestion, validation, processing, and delivery. We also discuss its benefits for security, memory and computational limits. Finally, we show that in testing, Databricks delivered runtimes that are 40 times faster than Stata on desktop for analysis through cluster auto-scaling and parallelization.

2. Cloud Storage & Computing for Central Banking

When working with larger data sets like NielsenIQ HomeScan with security concerns and scheduled deliveries, certain traditional practices have become outdated as storage and security needs become increasingly important. To eliminate some of these concerns, the Bank of Canada is piloting cloud-based solutions in Microsoft Azure.

Security

When dealing with protected data, traditionally, research groups at the Bank secured these data sets on various drives on the Bank's servers with special permissions and/or on the Bank's High Performance Computing (HPC) clusters. Researchers with approved access were given reading permissions to prevent accidental overwriting of raw data. With these permissions, researchers can copy out data to their local drives for analysis.

Without compromising security protocols, Azure provides researchers with single sign-on to access data, compute resources and other services[1]. In comparison, prior to the transition to Azure, researchers were required to maintain multiple credentials to be authenticated in multiple systems. In essence, by utilizing the Azure Active Directory (AD) service, Azure provides researchers with a hassle-free experience in accessing the data while protecting against security breaches.

[1] Murray, D. & Omondi, J. (12 May 2021). "What is single sign-on?" *Microsoft Azure*.

Since NielsenIQ HomeScan has a Protected-B security status, we put in place multiple measures to ensure only authorized personnel have the right access. Azure has a rich set of measures to ensure data security. On a high level, we utilize AD to authorize and authenticate researchers. For storage, we employ the Role-Based Access Control (RBAC) to control the access level to data containers, and the Access Control List (ACL) to folders and files under the containers. For other Azure services used by the pipeline, such as Azure Data Factory (ADF), Databricks, and SQL Databases, we use AD's single sign-on to manage researchers' access to data.

In the Azure Data Lake, we use Azure Gen2 Storage Accounts to store the various formats of NielsenIQ HomeScan data including CSV, Excel Workbooks, parquet, and delta[2]. Simultaneously, we use an Azure SQL Database to store tabular format data. Azure Gen2 Storage firewalls are configured to grant access to known IP addresses, in effect providing network security. The Azure Data Lake is deployed under an Azure Virtual Network (VNet)[3], which means the Bank has its own private network. Azure resources behind the VNet such as Azure SQL Database can communicate securely with each other through a virtual network service endpoint.

Memory Limits

When multiple researchers with approved access to protected data sets make copies locally in the pre-Azure environment, the memory required to allow multiple projects to progress simultaneously grows exponentially. This issue is resolved in Azure, since researchers can read from one central data storage location, so long as they have proper permissions to use analytical tools like Databricks. Moreover, within an Azure environment, researchers reduce data movement, which also reduces the rate at which the memory grows.

Computing Limits

Depending on the organization, requirements may exist with the use of compute resources for security reasons. If limits exist, researchers must create custom compute resources (called self-hosted runtime)[4]. A disadvantage of custom resources is the potential of memory limits. If the resources created for the organization are found to encounter memory issues, they can easily be scaled up in Azure without the need to buy physical hardware.

On the other hand, if security wasn't a concern, Azure has pretty much infinite computing power through Auto Resolve Runtime[5] that automatically scales the memory for users as required to complete processing.

[2] Brown, L. & McCready, M. (26 January 2022). "Delta Lake guide." *Azure Databricks*.

[3] Dwivedi, K., Berry, D. & Buck, A. (1 Feb 2022). "What is Azure Virtual Network?" *Microsoft Azure*.

[4] Li, L. & Burchel, J. (20 Jan 2022). "Create self-hosted integration runtime." *Microsoft Azure*.

[5] Li, L., Huff, A. & Burchel, J. (21 December 2021). "Integration runtime." *Microsoft Azure*.

Analytical Tools

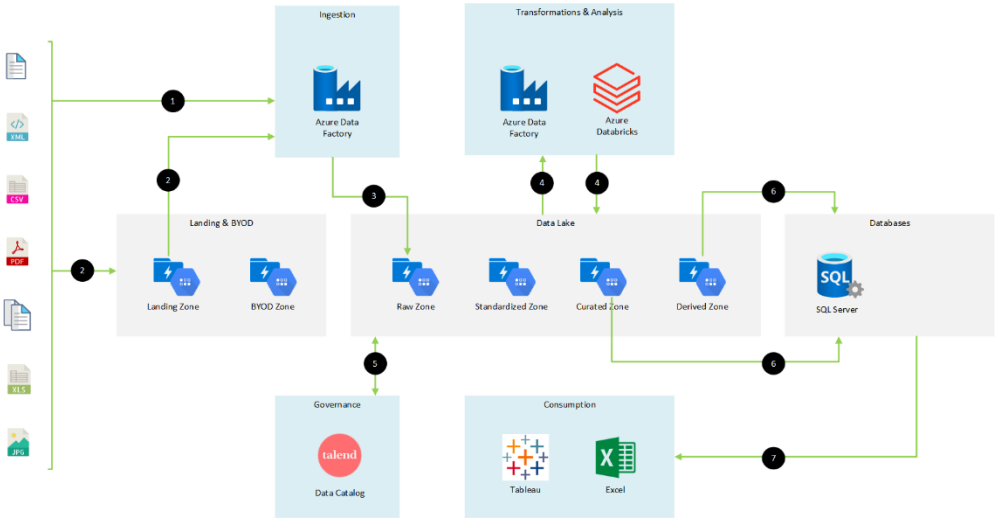
The Data Lake enhances researchers' analytical work by allowing their existing tools such as Stata, Matlab, Tableau and Python to connect to the database using JDBC[6] or ODBC[7] connectors. For some tools such as Python and Tableau, the Data Lake also allows researchers direct access to data under Azure Gen2 Storage Accounts which does not require connectors. In addition, researchers are also given access to Azure Databricks[8] notebooks to perform analysis using the highly scalable Databricks clusters. Section 4 describes the significant performance gains when performing analysis using Databricks notebooks.

3. Data Lake Pipeline

NielsenIQ HomeScan Pipeline

The design and implementation of the pipeline follow the Bank's Azure Data Lake architecture, as illustrated in **Figure 1**. The pipeline reflects the whole process of data flow that begins with ingestion, undergoes transformation, and concludes with loading for analysis. Data for each phase of the process is stored in different zones to provide data availability, traceability, and consistency to the highest level. ADF is used to orchestrate the end-to-end pipeline run.

Figure 1: Bank of Canada Data Lake Architecture



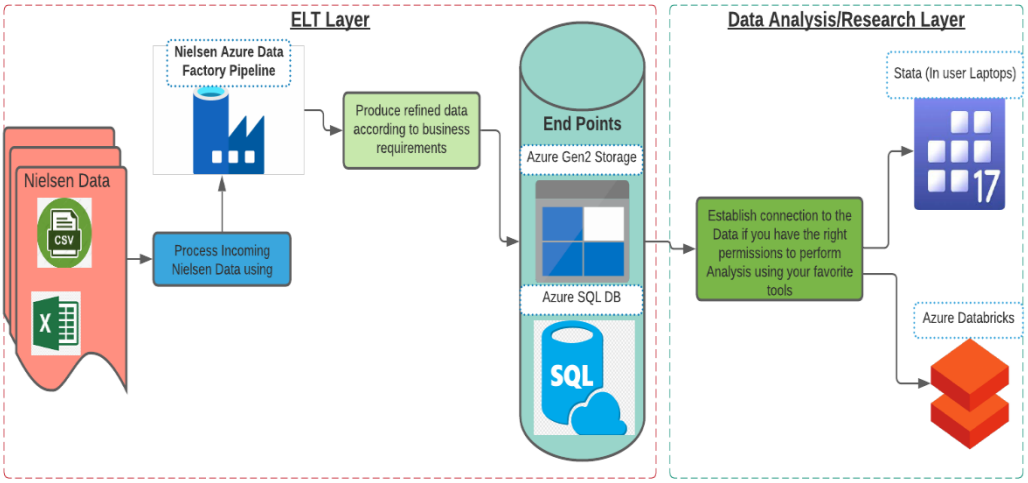
[6] Engel, D., Roth, J. & Sharkey, K. (8 December 2021). "Download Microsoft JDBC Driver for SQL Server." *Microsoft SQL*.

[7] Engel, D., Roth, J. & Sharkey, K. (2 November 2021). "Download ODBC Driver for SQL Server." *Microsoft SQL*.

[8] McCready, M. & Brown, L. (27 May 2021). "What is Azure Databricks?" *Microsoft Azure*.

Figure 2 below is a workflow for NielsenIQ HomeScan. The workflow starts from receiving data from the vendor, then processing and refining the data, and ends with delivery to researchers using various endpoints.

Figure 2: NielsenIQ HomeScan workflow

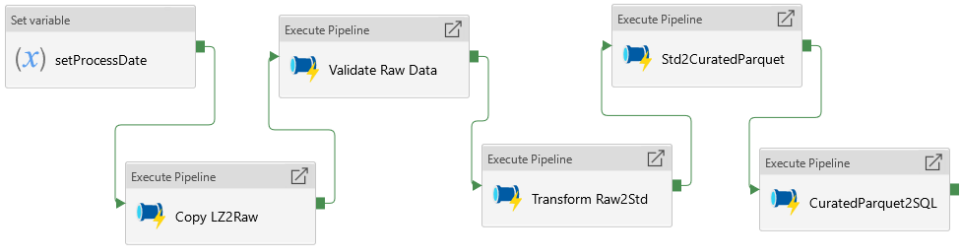


Notable Features of the Pipeline

Modularity

Figure 3 below illustrates the construct of the main pipeline, consisting of five highly modularized sub-pipelines. Each of the sub-pipelines could be run separately or jointly, providing greater flexibility and robustness of handling different use cases.

Figure 3: Nielsen Main pipeline



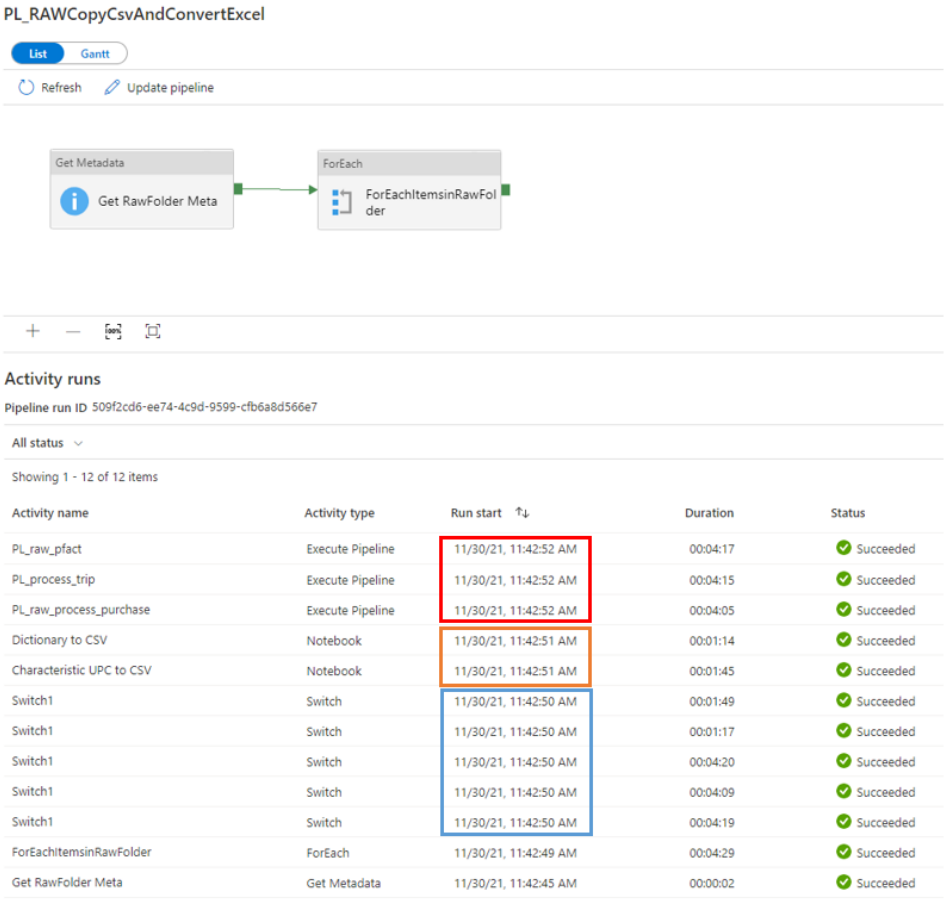
The pipeline aims to automate the process as much as possible, however we are also mindful of providing utilities to researchers to deal with actual cases. For example, in the "SetProcessDate" module, researchers can ask to process files from dates other than the default date by changing the process date parameter. Researchers may also want to pause an in-progress pipeline, then pick up later without losing any previous output. High modularity empowers researchers with such flexibility.

Performance Gains by Parallel Processing

The runtime for this pipeline has been reduced to a fraction of the pre-Azure runtime on a HPC server. These performance gains are obtained thanks to the parallel processing capability built in ADF and Databricks.

Using the "PL_CopyLZ2Raw" sub pipeline as an example in **Figure 4**, the distributed nature of Azure Data Lake storage, coupled with the "ForEach" construct in the ADF, enable the copying activities to be performed in parallel.

Figure 4: Parallel tasks processing



Another major contributor to performance gains is the use of Databricks clusters in handling the file loading and consequent transformations. Databricks builds on Apache Spark[9], a well-known project dedicated to big data and distributed computing.

[9] Zaharia, M. (19 Nov 2018). "What is Apache Spark?" *Databricks*.

Data quality enhancements via automated data validation

Data quality assurance is vital to the quality of research produced. In this pipeline, we embed procedures to improve data quality. Embedded procedures have the following benefits:

- Automated data validation for future data deliveries, thus ensuring we do not corrupt existing data with new invalid data
- Reliable and consistent data validation compared to manual intervention
- Fast runtimes, thus providing almost instantaneous feedback to the data vendor

We use a program to validate incoming data according to the vendor's technical specifications. In the past, we relied mainly on manual efforts to spot potential errors.

The program validates business logic, which are difficult to identify manually. During development, this program helped the Data Lake team identify loss of precision in unique numeric identifiers, as well as business logic errors. For example, we found inconsistencies in the translation from the week a shopping trip took place to the corresponding trip month. Business logic errors similar to the previous example created inaccuracies in the analysis of the aggregated data, especially when weeks in December were categorized in the aggregated data as November or January. This shifting of weeks skewed seasonal shopping trends prior to correction.

Maintaining data security during the data validation process

When the data fails any specified logic, a unique error message is logged to Azure Log Analytics and appropriate Bank personnel are alerted. However, Log Analytics is designed to be accessed by everyone at the Bank, which can be a problem if the data is protected.

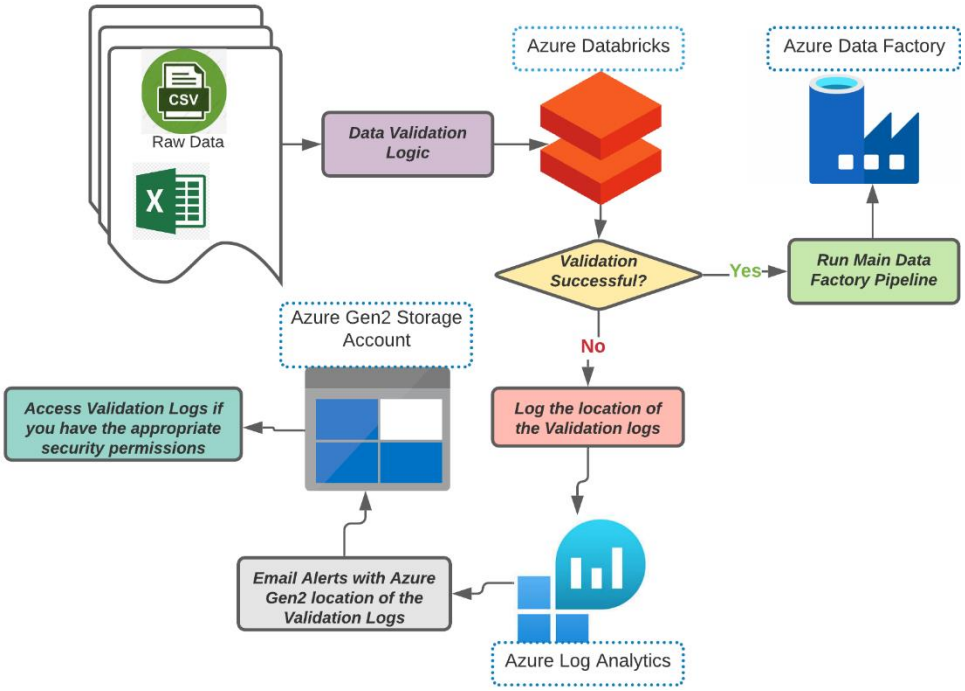
To ensure protected data is not accessed by unauthorized users, the data validation failures are logged to Log Analytics describing the *location* of the detailed logs in Azure Gen2 Storage rather than the *content* of failures. In Azure Gen2 Storage, data is secured using AD groups. Therefore, someone with the correct credentials can login and see the detailed data validation failure and share it with the vendor for resolution. **Figure 5** below describes this process.

4. Analysis & Performance

Loading data into the Data Lake

Prior to the Data Lake pipeline development, new data deliveries required a researcher-run data cleaning procedure. This involved loading raw Excel Workbook data into Stata and merging with household-level sampling weights, data dictionary files, as well as household-level monthly spending summary files. Given the size of the data set, this procedure took upwards of 3-4 hours running on the Bank's HPC clusters at each delivery, even after runtime reduction procedures such as transforming all Workbook files to CSV. The NielsenIQ HomeScan pipeline only requires upwards of 20 minutes to load new data upon delivery.

Figure 5: Maintaining security during data validation



Analytical performance

The benefit of data storage in the cloud is the ease of analysis using analytical tools such as Databricks. In **Table 1**, we summarize the performance improvements of Databricks compared to two alternative methods on a NielsenIQ HomeScan data set with over 61.5 million observations. In the time it requires to run the same analysis *once* in Stata on desktop, Databricks could run it more than 40 times.

Table 1: Comparison of analytical runtimes, seconds (s)

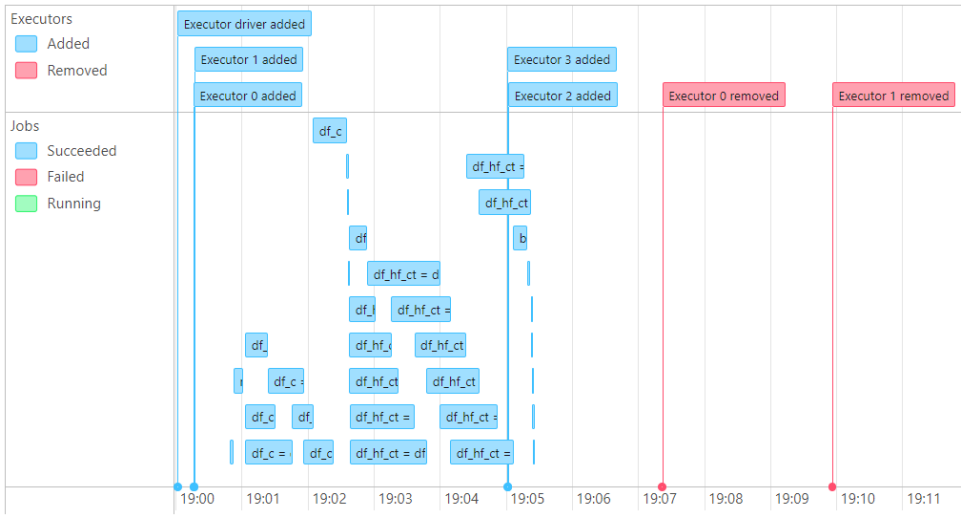
	Loading data set	Data cleaning	OLS regression	Total runtime
Stata (desktop)	9,800 –12,897	401	250	10,451 – 13,548
Stata (HPC)	9,800 –12,897	224	110	10,134 – 13,231
Databricks*	13	137	107	257

*At peak runtime, 64GB and 16 cores were allocated to this job. Highlighted are fastest runtimes in each category.

Databricks notebooks are run on a cluster that consists of a driver node and different numbers of worker nodes depending on the workload. The driver node and each of the worker nodes is configured with 14GB memory and 4 cores. At peak time, the cluster auto-scaled to 4 worker nodes. **Figure 6** visually illustrates how the analysis described in **Table 1** is completed in Databricks notebooks. We can see that

the analysis begins around 19:01 and completes in a little over 4 minutes as reported. The cluster parallelizes certain processes automatically to reduce runtimes. Then, once the Databricks notebook goes stale around 19:05, the cluster automatically removes worker nodes as they are no longer required by this specific notebook we are running.

Figure 6: Parallel computing and cluster auto-scaling in Databricks



5. Conclusion

Large organizations like central banks can adopt cloud-based technology to significantly improve computational capabilities while ensuring security and reducing costs. Cloud-based Data Lakes allow for an integrated and automated data pipeline that can securely store and back-up data of multiple formats, structured or unstructured. Analytical tools like Databricks are integrated with Azure Data Lakes, reducing data movement within an organization while delivering significant performance gains.

In this paper, we discussed our implementation of an Azure Data Lake for the Canadian NielsenIQ HomeScan data set. We discussed an automated data pipeline and its benefits for security, memory and computational limits, and assuring data quality. Finally, we showed that, through cluster auto-scaling and parallelization, Databricks delivered runtimes that are 40 times faster than Stata on desktop for analysis during testing.

BANK OF CANADA
BANQUE DU CANADA

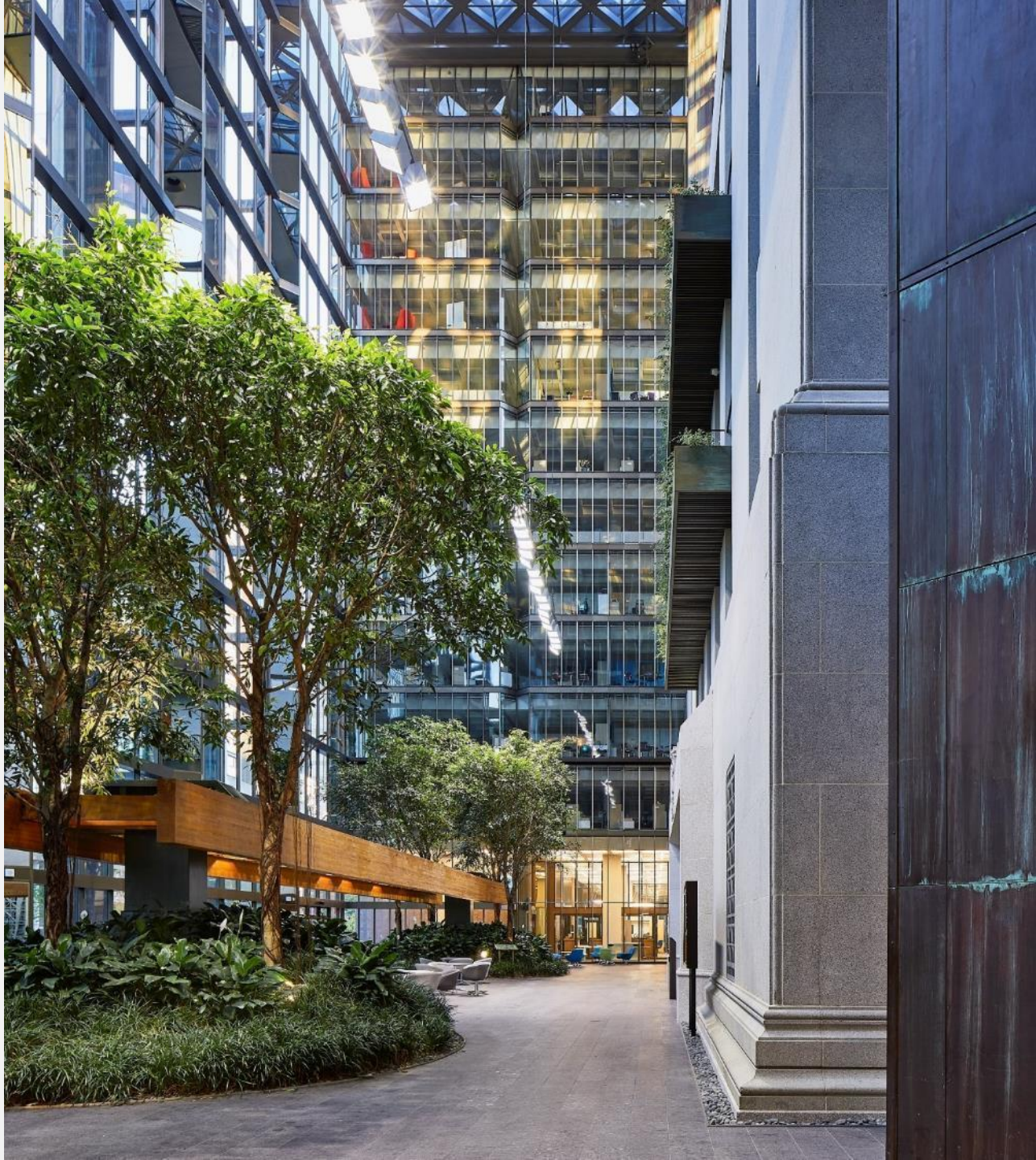
IFC-BANK OF ITALY DATA SCIENCE IN CENTRAL BANKING WORKSHOP
16 FEBRUARY 2022/16 FÉVRIER 2022

Swimming in the Data Lake

An application to NielsenIQ HomeScan

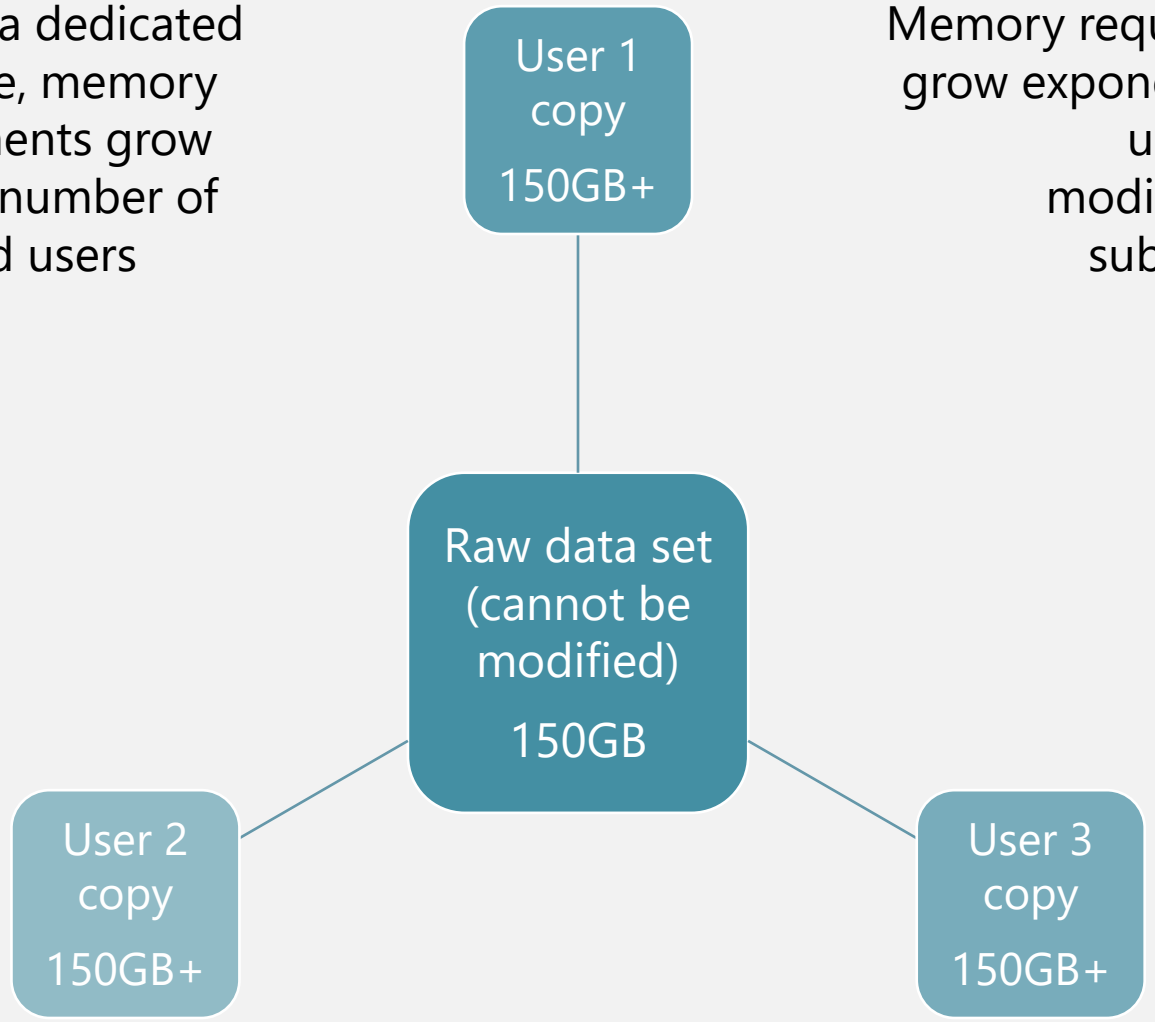
Minnie H. Cui

JOINT WITH BOTLHALE MOSWEU & GENE (FA GUI) JIANG



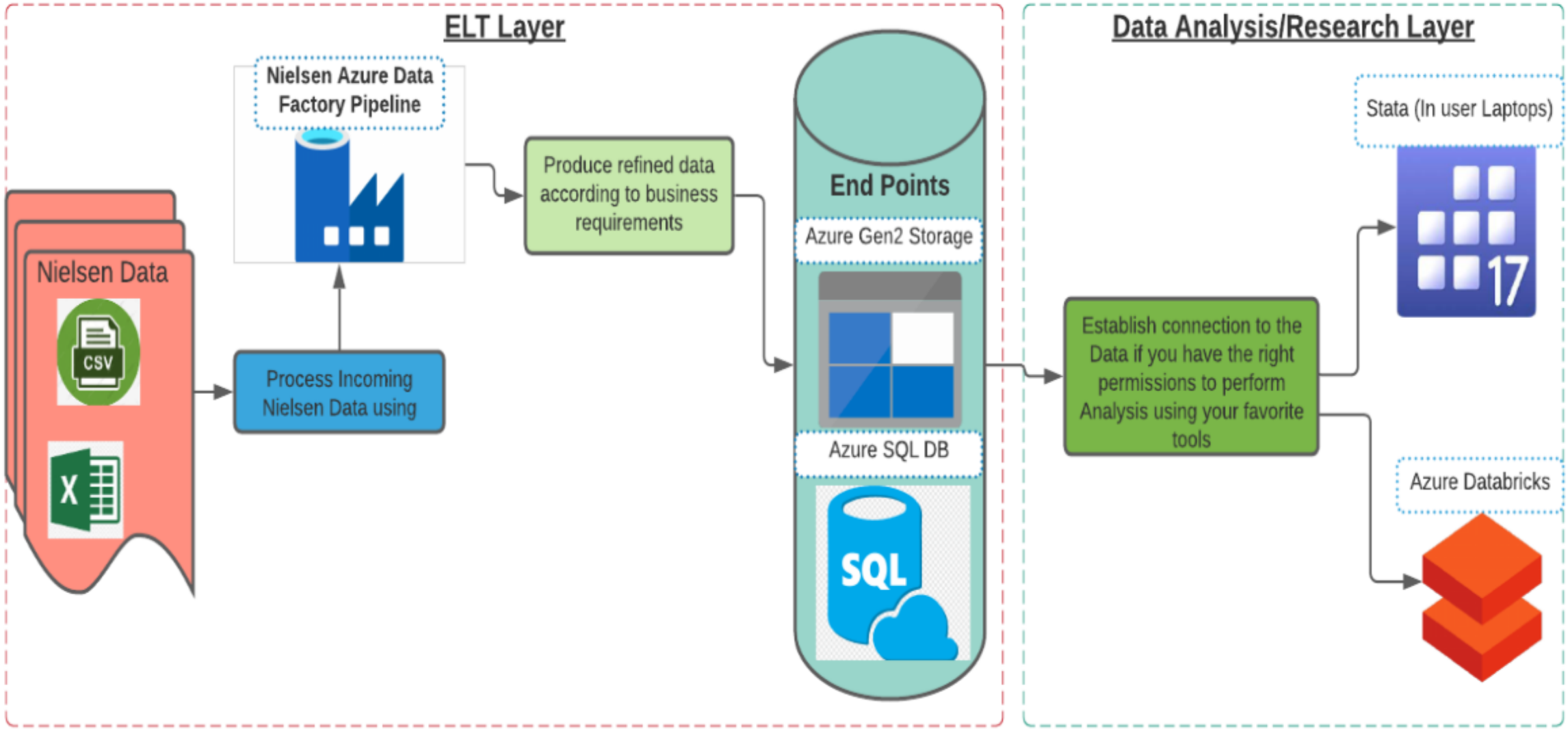
Pre-cloud protected data storage at BoC

Without a dedicated data base, memory requirements grow with the number of approved users

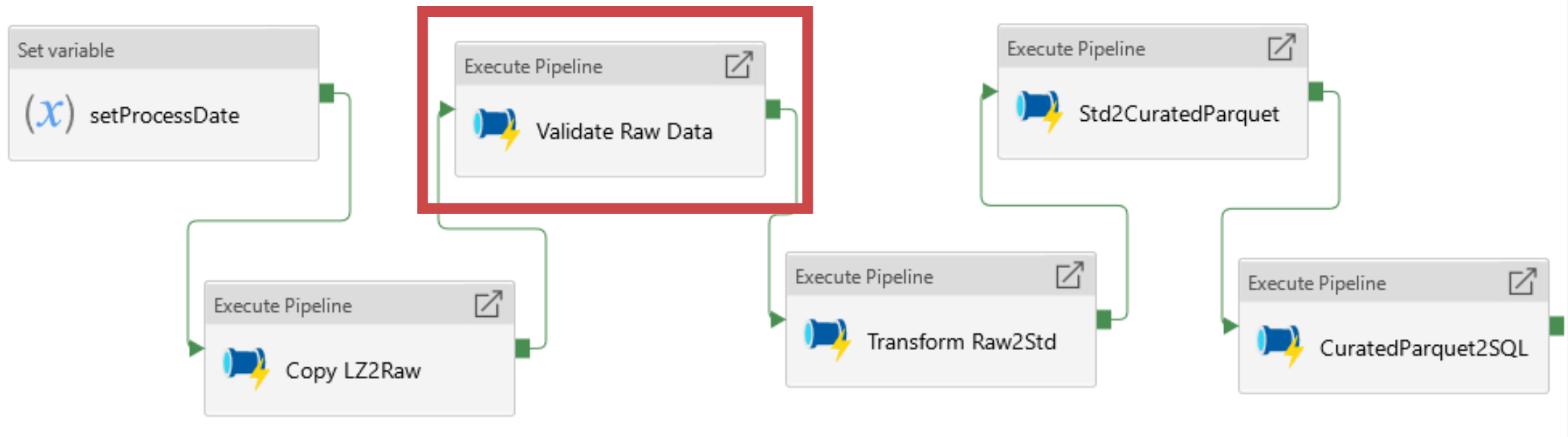


Memory required can grow exponentially if users save modifications, subsets, etc.

NielsenIQ HomeScan workflow



NielsenIQ HomeScan pipeline



Benefits of automated validation on data quality

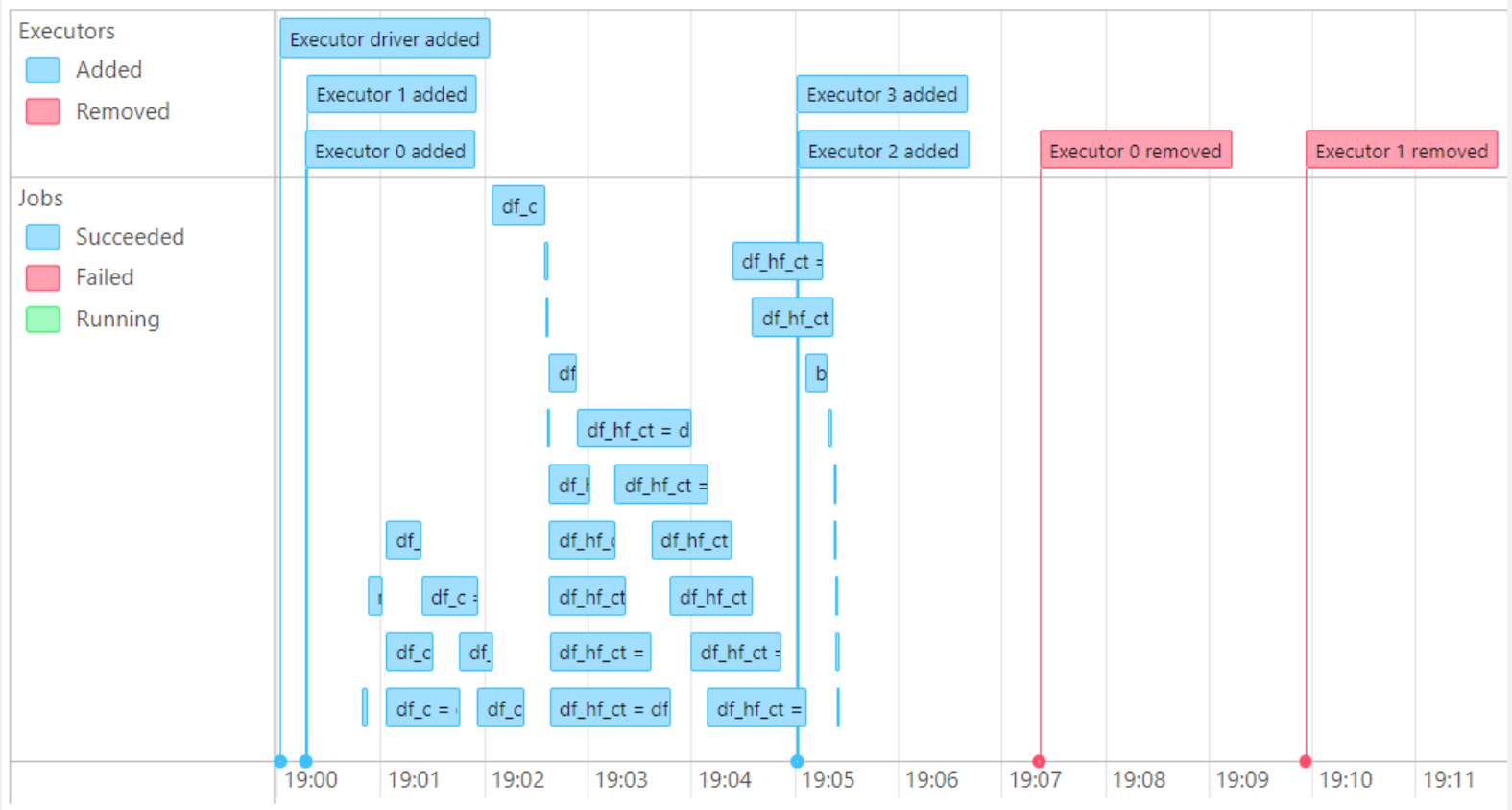
- Automated data validation for future data deliveries thus ensuring we do not corrupt existing data with new invalid data
- Reliable and consistent data validation compared to manual intervention
- Runs fast providing almost instantaneous feedback to the data provider

Comparison of analytical runtimes, in seconds (s)

	Loading data set	Data cleaning	OLS regression	Total runtime
Stata (desktop)	9,800 – 12,897	401	250	10,451 – 13,548
Stata (HPC)	9,800 – 12,897	224	110	10,134 – 13,231
Databricks*	13	137	107	257

*At peak runtime, 64GB and 16 cores were allocated to this job. Highlighted are fastest runtimes in each category.

Parallel computing and cluster auto-scaling in Databricks



Certain processes are automatically run in parallel to reduce runtime

Cluster automatically removes worker nodes when the analysis is complete

Ensuring security

- Without compromising security protocols, Azure provides users with single sign-on access to data
- Given HomeScan's Protected-B status, we use
  - Azure Active Directory (AD) to authorize and authenticate users
  - Role-based access control (RBAC) to control the access level to data containers
  - Access control list (ACL) to folders and files under the containers

Ongoing questions about security

- BoE has expressed concern with “secretive” nature of cloud computing providers
- Is the data in the cloud secure from unauthorized access, for instance, from the provider?
  - Azure Data Lake implementation at the BoC is behind an Azure Virtual Network (VNet)
- What is the risk that multiple banks, for instance, can be affected by cyber attacks or service outages?
- These issues should be discussed more before a large scale implementation…


Questions?