Technical Validation Success Plan

Databricks Background

Founded by the original creators of Apache Spark™, Delta Lake and MLflow, Databricks brings together data engineering, science and analytics on an open, unified platform so data teams can collaborate and innovate faster. More than 5000 organizations worldwide rely on Databricks as a unified platform for massive-scale data engineering, collaborative data science, full-lifecycle machine learning and business analytics. With offices on four continents and hundreds of global partners, Databricks is on a mission to help data teams solve the world's toughest problems.

The company was founded in 2013 and is headquartered at San Francisco, California.

PayU Background

PayU is a global online payment service provider, offering secure and convenient payment solutions for businesses and consumers. PayU enables merchants to accept payments through various channels, including credit/debit cards, bank transfers, and digital wallets. Their platform provides advanced fraud prevention tools and seamless integration options for e-commerce websites and mobile apps. PayU's commitment to innovation drives their development of new technologies and features to enhance the payment experience. Trusted by millions of users worldwide, PayU facilitates fast and reliable transactions, empowering businesses to grow and thrive in the digital economy.

Engagement Overview

PayU Payments is evaluating Databricks Lakehouse Platform on AWS to replace their existing reporting Platform components using redshift and solve the associated challenges for their reporting workloads. PayU and Databricks have agreed to pursue a joint Proof of Concept (POC) on the Databricks platform and will be working together to complete this technical validation. With Databricks, PayU can accelerate innovation by unifying data engineering, data science and business analytics in one platform.

This document outlines the current state, future state, scope of engagement, roles & responsibilities and evaluation criteria for the technical validation.

Databricks will provide the following:

- A technical contact to help with the deployment of Databricks on AWS Cloud Platform.
- Databricks subject matter experts to assist with the POC / evaluation of the Databricks platform.

PayU will provide:

- A Site Reliability Engineer (SRE) or DevOps engineer to help deploy Databricks and integrate with their AWS infrastructure.
- Data Engineers and Data Analysts to evaluate Databricks.
- Updates to the POC evaluation criteria and results/metrics of the evaluation as needed.

Upon completion of the technical validation success criteria, PayU and Databricks will co-present results to the PayU executive team and both parties will consider a mutually beneficial commercial agreement.

Key Contacts

PayU	Databricks
•	• Jeremiah Peter- Account Executive • Sumaer Bahl- Sr. Manager, Digital Native Business • Raghuraj Vemula - Solutions Architect

Groups & Teams Involved in Evaluation

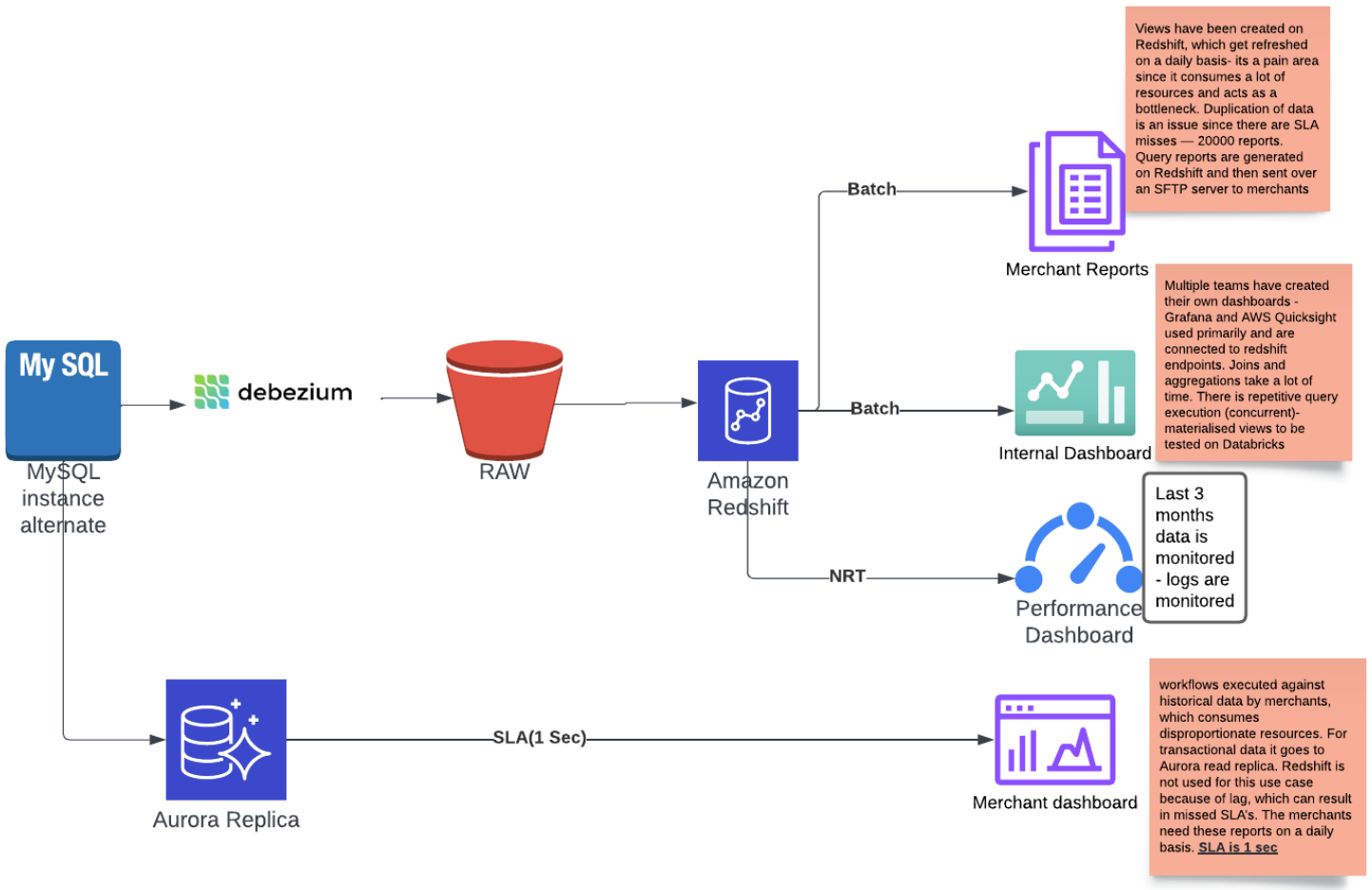
1. PayU Cloud Infrastructure / Admin Team
2. PayU Infosec / Data Security Team
3. PayU Data Platform Engineering Team

Technical Validation Timeline

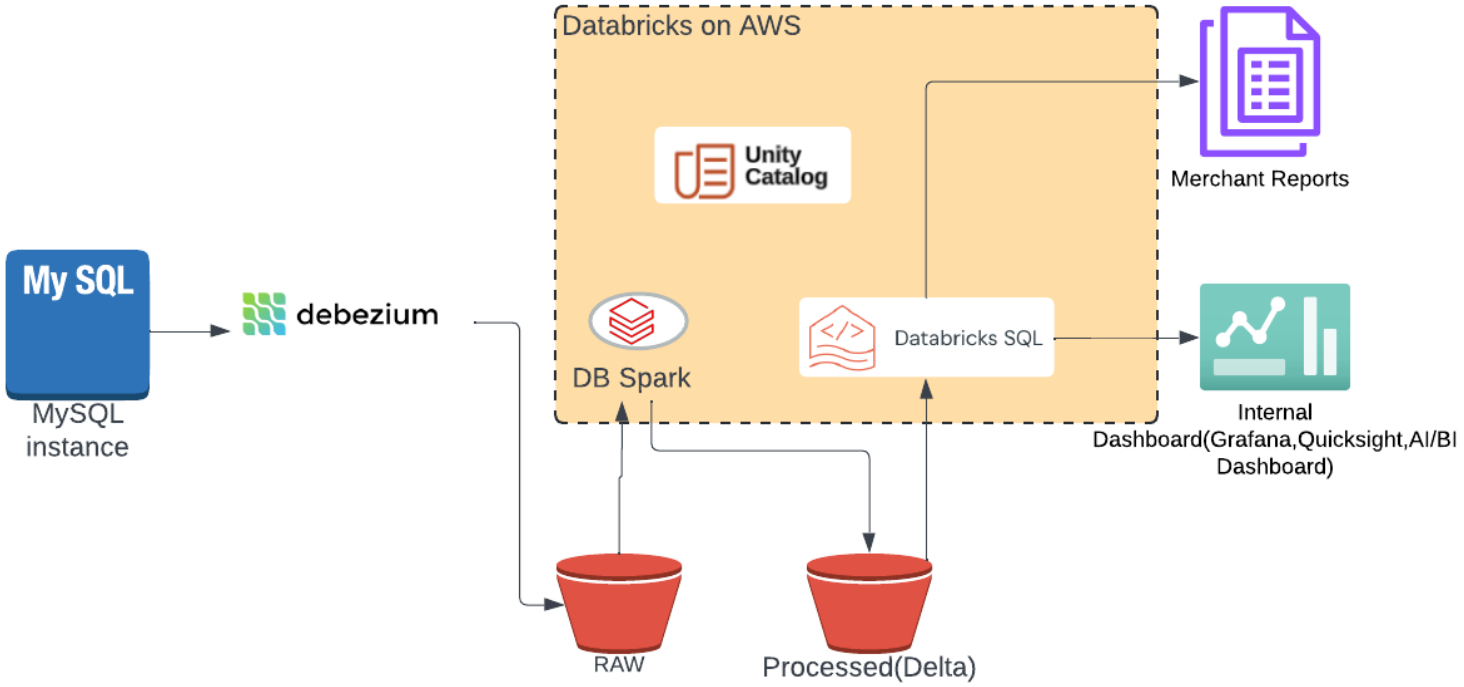
The evaluation will commence and conclude at mutually agreed start and end dates mentioned below.

Start Date	05/12/2024
End Date	

Current State:


My SQL
MySQL instance alternate
debezium
RAW
Amazon Redshift
Batch
Merchant Reports
Views have been created on Redshift, which get refreshed on a daily basis- its a pain area since it consumes a lot of resources and acts as a bottleneck. Duplication of data is an issue since there are SLA misses — 20000 reports. Query reports are generated on Redshift and then sent over an SFTP server to merchants
Batch
Internal Dashboard
Multiple teams have created their own dashboards - Grafana and AWS Quicksight used primarily and are connected to redshift endpoints. Joins and aggregations take a lot of time. There is repetitive query execution (concurrent)- materialised views to be tested on Databricks
NRT
Performance Dashboard
Last 3 months data is monitored - logs are monitored
Aurora Replica
SLA(1 Sec)
Merchant dashboard
workflows executed against historical data by merchants, which consumes disproportionate resources. For transactional data it goes to Aurora read replica. Redshift is not used for this use case because of lag, which can result in missed SLA's. The merchants need these reports on a daily basis. SLA is 1 sec

Databricks Flow(POC):



Evaluation Criteria

Objective	Business Value	Measures of Success	[Company] Owners	Status
Validate Governance	Enhanced security and governance implementation	To be able to access control across different assets like tables, files row/Columnar access PII Data Masking		
Data Analytics & reporting	Data Availability	Specific report / query performance (Testbed of approx 5-10 queries) ● 30% complex ● 40% medium ● 30% simple 5-10 small queries - Meet SLA of < 1min & SLA of < 5 mins 3 long running queries - Compare to perform better against current execution Price to performance ratio		

Separate ETL workload from Redshift to Spark on Databricks	Decouple storage and compute to a future proof architecture	Reduced cost of ETL vs Redshift End to end latency to meet current SLA		

Use Cases & Requirements for Evaluation

This section is to codify the use cases, and specific requirements, that PayU will be evaluating during the Technical Validation.

1. **Use Cases**

#	Use Case	Success Criteria
1	Merchant Reports	Better price to performance compared to current Redshift batch flow.
2	Internal Dashboards	Ease of use of AI/BI dashboards and TCO.

2. **Scope and Volume of data & tables**

#	Files / DB	Volume	History	Use cases	Size of history	True source
1	<sample - please fill it with originals>Orders	3M/day	6 Months	Batch	60 Million	My SQL

#	Files / DB	File Sizes	Filters	Keys
1	<sample - please fill it with originals>Orders		created_on, status	id, code, store_id, user_id, [ALL _id columns]

POC Results

Technical Results & Metrics

#	Use case	Requirement & SLA	SKUs used & Details	Outcome
1	DA Performance: Short running queries	• Test <10 short queries • SLA of < 1min & SLA of < 5 mins	DB SQL Classic (S)	
2	DA Performance: Long running queries	• Test critical 3 complex & long running queries • Compare against current performance	DB SQL Pro (M) / Databricks Jobs (with smaller config)	
3	DG: Validate Governance feature	• To be able to access control across different assets like tables, files	Unity Catalog	

Proof of Concept (POC) - Technical Evaluation Timeline and Tasks

Area	Testing Requirement	Owner	Start	End	Status

E2 Account Creation					
	• Cross Account IAM Role • Setting up root bucket • Workspace Creation	PayU Team	1st week		
Administration					
	• Launch clusters • Cluster sizing and administration • Best practices • Databricks Rest APIs • Admin Console	PayU Team	1st week		
Connecting to Data Sources					
	• Mount S3 buckets using DBFS • AWS Keys • IAM Roles • IAM Credential Passthrough • Hive Metastore Integration	PayU Team	1st week		
Data Access and Security					
	• Token based authentication • Databricks CLI • Table Access Controls • Secrets API	PayU Team	1st week		
Data Ingestion and ETL					

	• Delta Auto Ingest • Delta Auto Optimize • Delta Optimize • Delta Z Ordering • Delta Lake and Delta Engine Guide	PayU Team	1st week		

POC - Issues Tracker

#	Question/Issue	Reported By	Date Reported	Status	Action/Resolution