



DESPLEGAMENT D'INFRAESTRUCTURA END TO END PER A LA GESTIÓ DE DADES

TREBALL DE FINAL DE MÀSTER

MÀSTER EN BIG DATA – LA SALLE URL

ALUMNE: POL GRÀCIA ESPELT

TUTOR: JOAN NAVARRO

CONTINGUTS

- Context
- Objectiu
- Abast
- Marc Teòric
- Proposta de solució
- Desenvolupament i implementació
- Cas d'estudi
- Conclusions

The background is a blue gradient. In the corners, there are white line art elements resembling circuit traces or neural network connections, with small circles at the end of the lines.

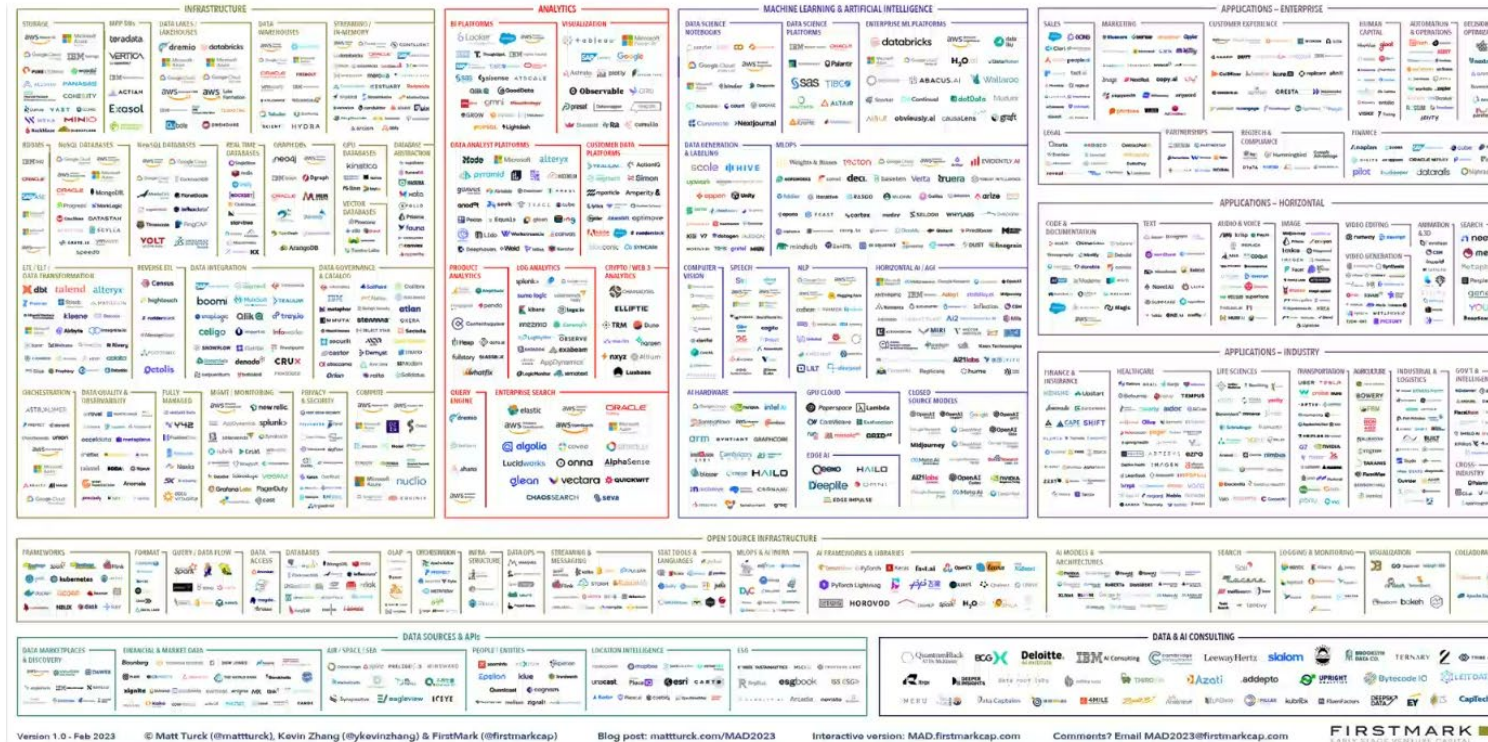
CONTEXT

LA REVOLUCIÓN DEL VALOR DE LAS DADES

- Companies *data-driven*
- Mejora constante de las arquitecturas
- Evolución del mercado de Servicios en Big Data



The 2023 ML, AI, and Data Landscape




MERCAT DEL BIG DATA

- Altament Competitiu
- En constant evolució
- Amb gran varietat
- Alt grau d'especialització

The background is a blue gradient. In the corners, there are decorative white line art elements resembling circuit boards or neural networks, with lines and small circles.

OBJECTIU



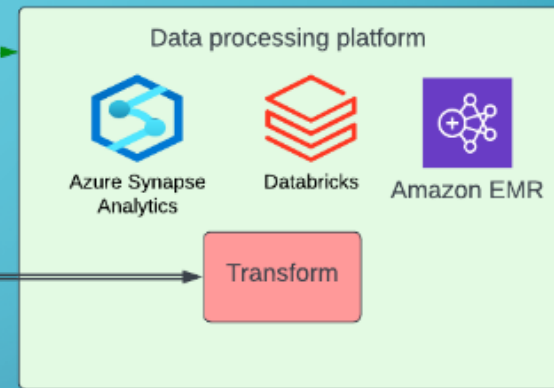
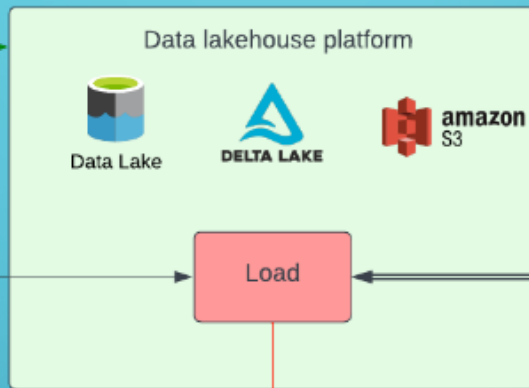
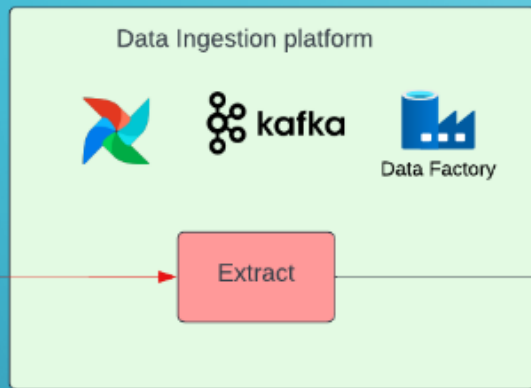
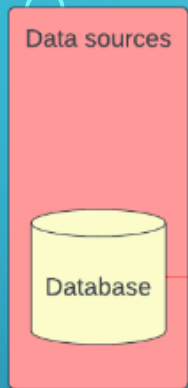
DESENVOLUPAR UN SISTEMA QUE DE MANERA AUTOMÀTICA DESPLEGUI TOTS ELS COMPONENTS NECESSARIS PER A LA CREACIÓ D'UNA ARQUITECTURA END-TO-END PER A LA GESTIÓ DE DADES, QUE ES COMPONGUI DE LES CAPES D'ADQUISICIÓ, EMMAGATZEMATGE, PROCESSAMENT I CONSUM QUE SIGUI INDEPENDENT DELS SEUS ORÍGENS, CONTINGUT I VOLUM I COMPLEIXI AMB ELS REQUISITS D'ACTUALITAT, EFICIÈNCIA I SEGURETAT.

ES PRETÉN PAQUETITZAR UNA SOLUCIÓ D'ARQUITECTURA GENÈRICA QUE FACILITI LA IMPLEMENTACIÓ DE PROJECTES DE BIG DATA I PERMETI A LES ORGANITZACIONS CONCENTRAR-SE EN L'EXTRACCIÓ DEL VALOR DE LES SEVES DADES, MINIMITZANT LA COMPLEXITAT DEL DESPLEGAMENT I LA ELECCIÓ DE COMPONENTS.



The background is a blue gradient. In the corners, there are white line-art illustrations of circuit boards. These include straight lines, right-angle turns, and small circles representing solder points or vias. The lines vary in thickness and are distributed across all four corners, with a higher density on the left side.

ABAST



The background is a blue gradient. In the corners, there are white line-art illustrations of circuit boards or data paths, with lines connecting to small circles.

CAVES DE L'ARQUITECTURA BIG DATA

ES BUSCA ANALITZAR EL MERCAT I BUSCAR SOLUCIONS CONCRETES PER CADA CAPA QUE COMPLEIXIN AMB ELS OBJECTIUS DEL PROJECTE.

CAPA D'INGESTIÓ I RECOPILOCACIÓ

IMPORTÀNCIA

Capa que s'encarrega de la recollida i connexió entre les fonts de dades i els entorns d'analítica.

DESAFIAMENTS

- Escalabilitat
- Latència
- Manteniment de la integritat

MERCAT

- AWS Glue
- Azure Data Factory
- Apache Nifi

CAPA D'INFRAESTRUCTURA I EMMAGATZEMATGE

IMPORTÀNCIA

Capa que s'encarrega de l'aprovisionament de recursos computacionals, gestió del emmagatzematge i xarxa d'interconnexió.

DESAFIAMENTS

- Escalabilitat vertical i horitzontal
- Consistència i disponibilitat
- Seguretat
- Interoperabilitat

MERCAT

- AWS (s3)
- Azure (Blob storage)
- GCP (Google cloud storage)
- HDFS

CAPA DE PROCESSAMENT I TRANSFORMACIÓ

IMPORTÀNCIA

Capa que s'encarrega de la manipulació i anàlisi de les dades emmagatzemats a la capa anterior.

DESAFIAMENTS

- Gestió batch i streaming
- Concurrència
- Tolerància a errors

MERCAT

- Apache Spark
- Databricks
- Amazon EMR
- Azure HDInsight

CAPA D'ORQUESTRACIÓ I AUTOMATITZACIÓ

IMPORTÀNCIA

Capa que s'encarrega de la coordinació i automatització de tasques i flux de treballs en el sistema. És un controlador centralitzat entre les capes d'ingestió, emmagatzemant i processament.

DESAFIAMENTS

- Complexitat en la coordinació
- Resiliència
- Elasticitat

MERCAT

- Apache Airflow
- Aws Step Functions
- Azure Logic Apps
- Kubernetes

CAPA D'ANÀLISI I CONSULTA

IMPORTÀNCIA

Capa que s'encarrega de la extracció de valor en les dades en forma de gràfiques, agregacions i informes.

DESAFIAMENTS

- Seguretat
- Interactivitat
- Integració

MERCAT

- Power BI
- QlikView
- Sisense

CAPA DE SEGURETAT I COMPLIMENT

IMPORTÀNCIA

Capa que s'encarrega de la protecció de la integritat, disponibilitat i confidencialitat de les dades. També aborda temes en aspectes regulatoris i legals.

DESAFIAMENTS

- GDPR o regulació
- Escalabilitat
- Control d'accés

MERCAT

- Okta
- MPI
- Varonis

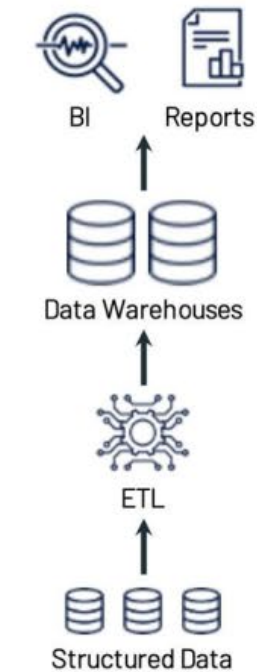


ARQUITECTURA

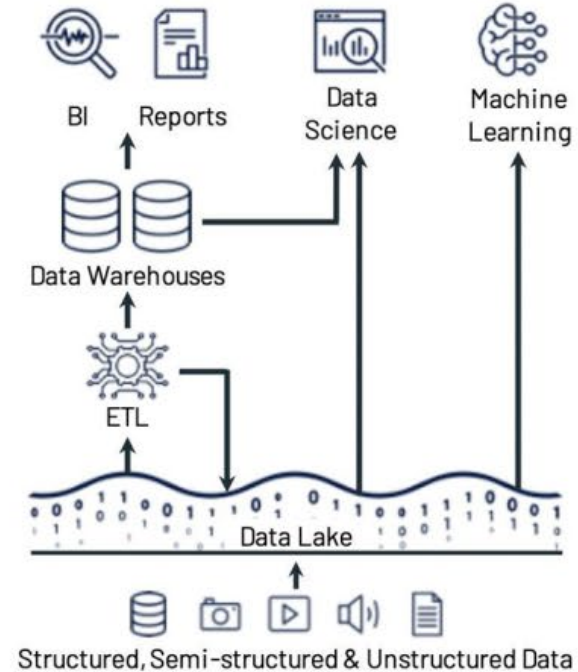
ES BUSCA INTEGRAR ELS COMPONENTS EN UNA ARQUITECTURA QUE COMPLEIXI ELS OBJECTIUS DEL PROJECTE.



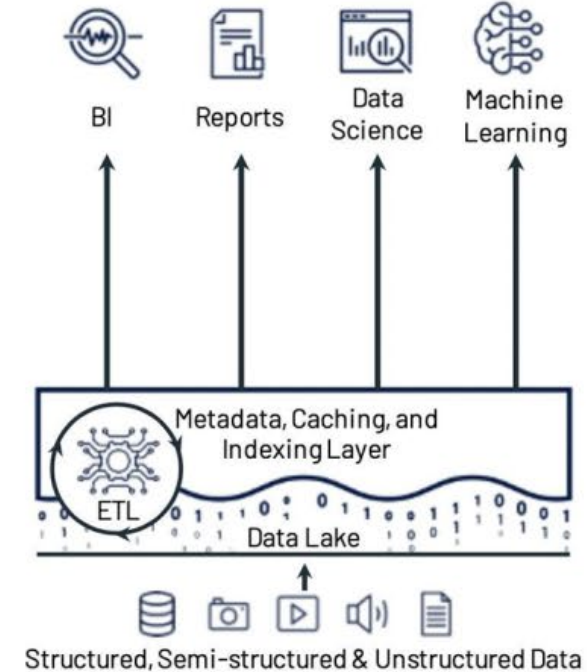
Evolution of Big Data Platform Architecture



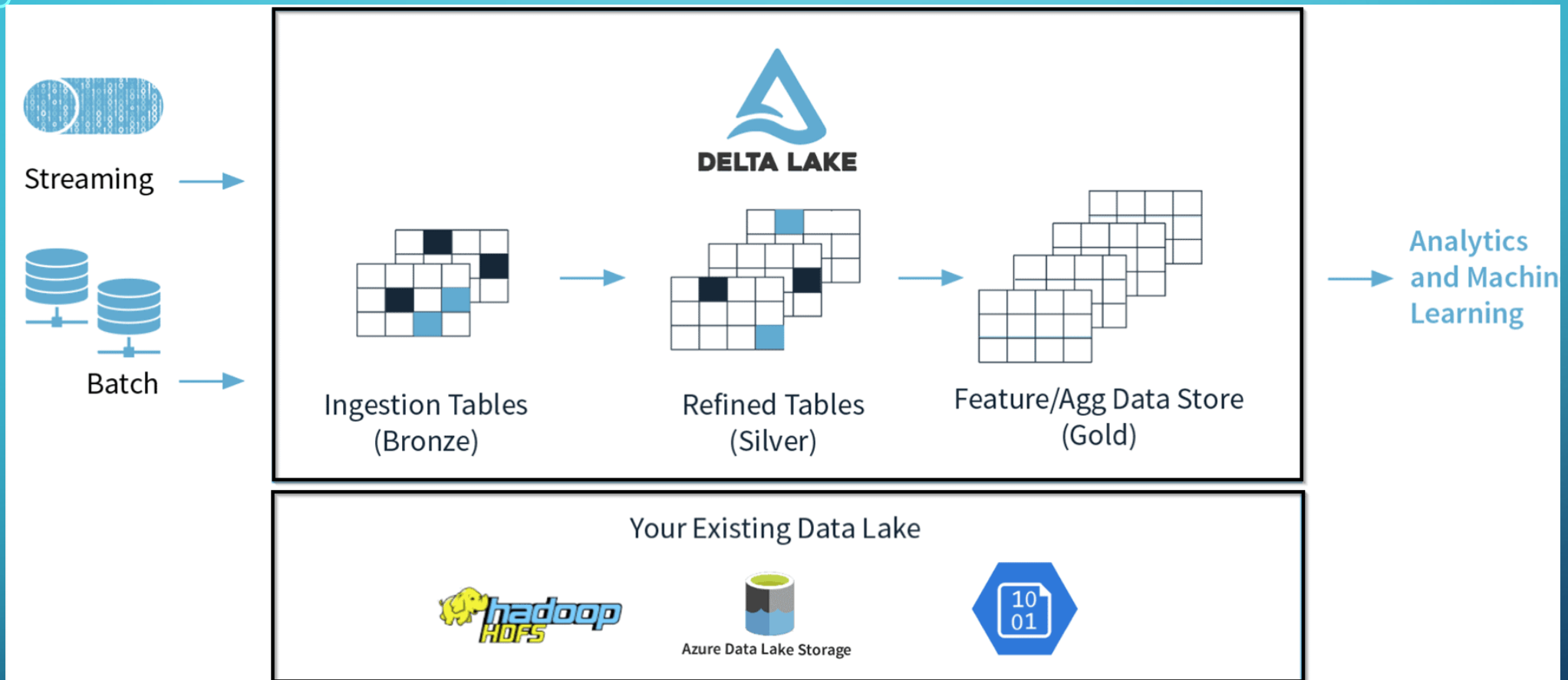
Data Warehouse Architecture



Data Lake-Warehouse Architecture



Lakehouse Architecture



- Compensió columnar basada en parquet.
- Asegura transaccions Atòmiques, consistents, aïllables i durables.
- Manté un log de transaccions que permet traçar el linatge i historial dels fitxers.

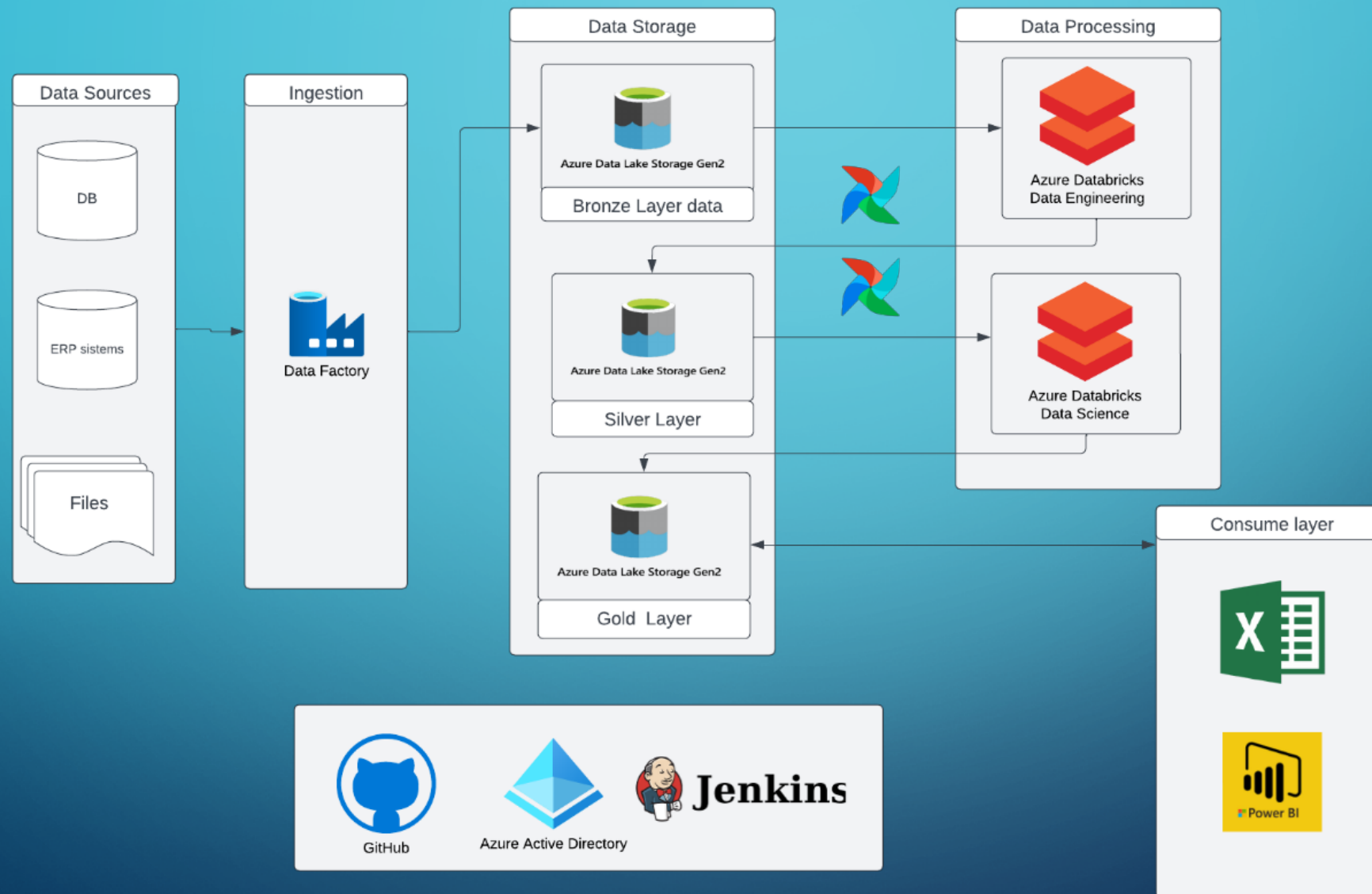


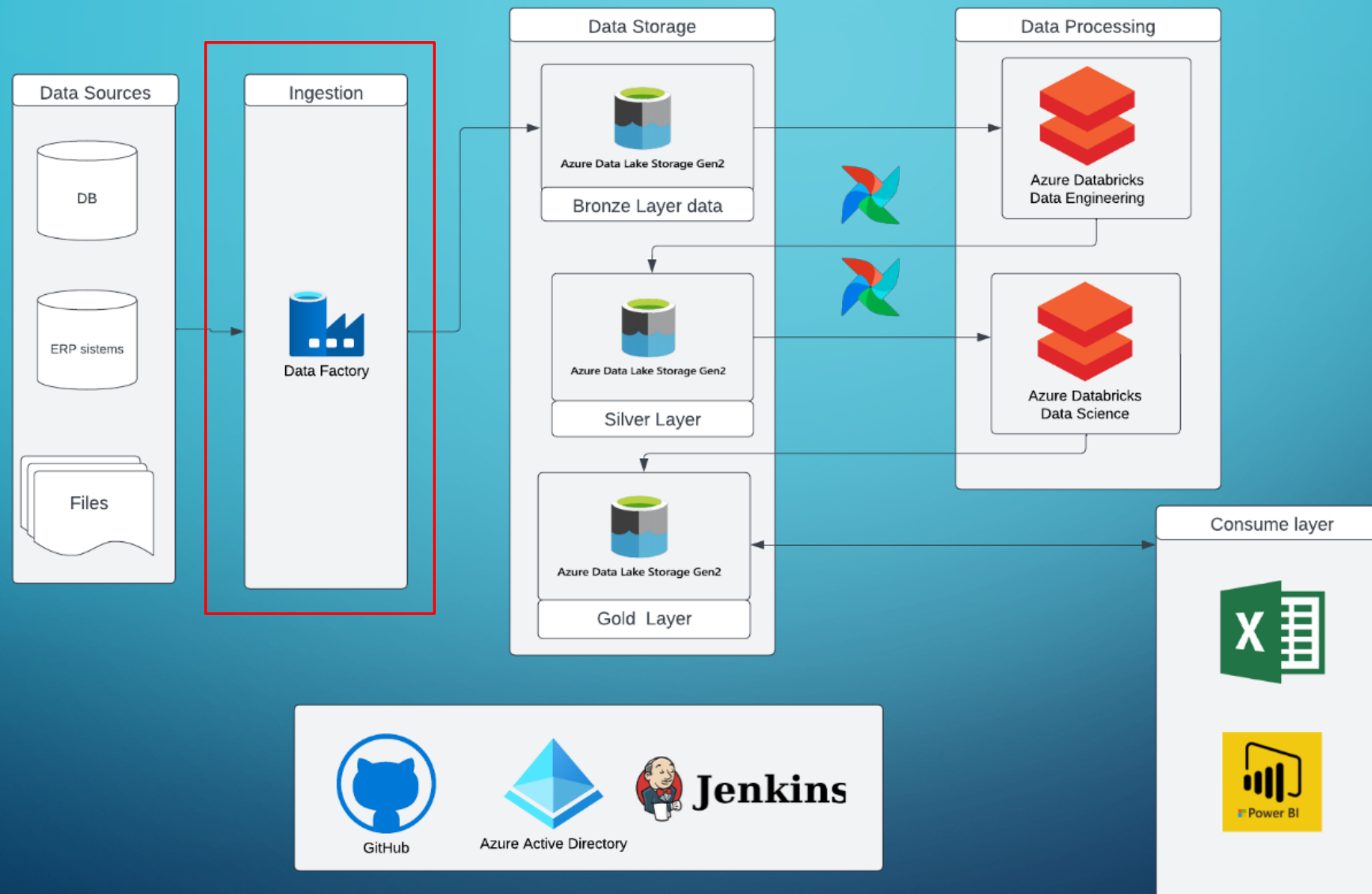
PROPOSTA DE SOLUCIÓ

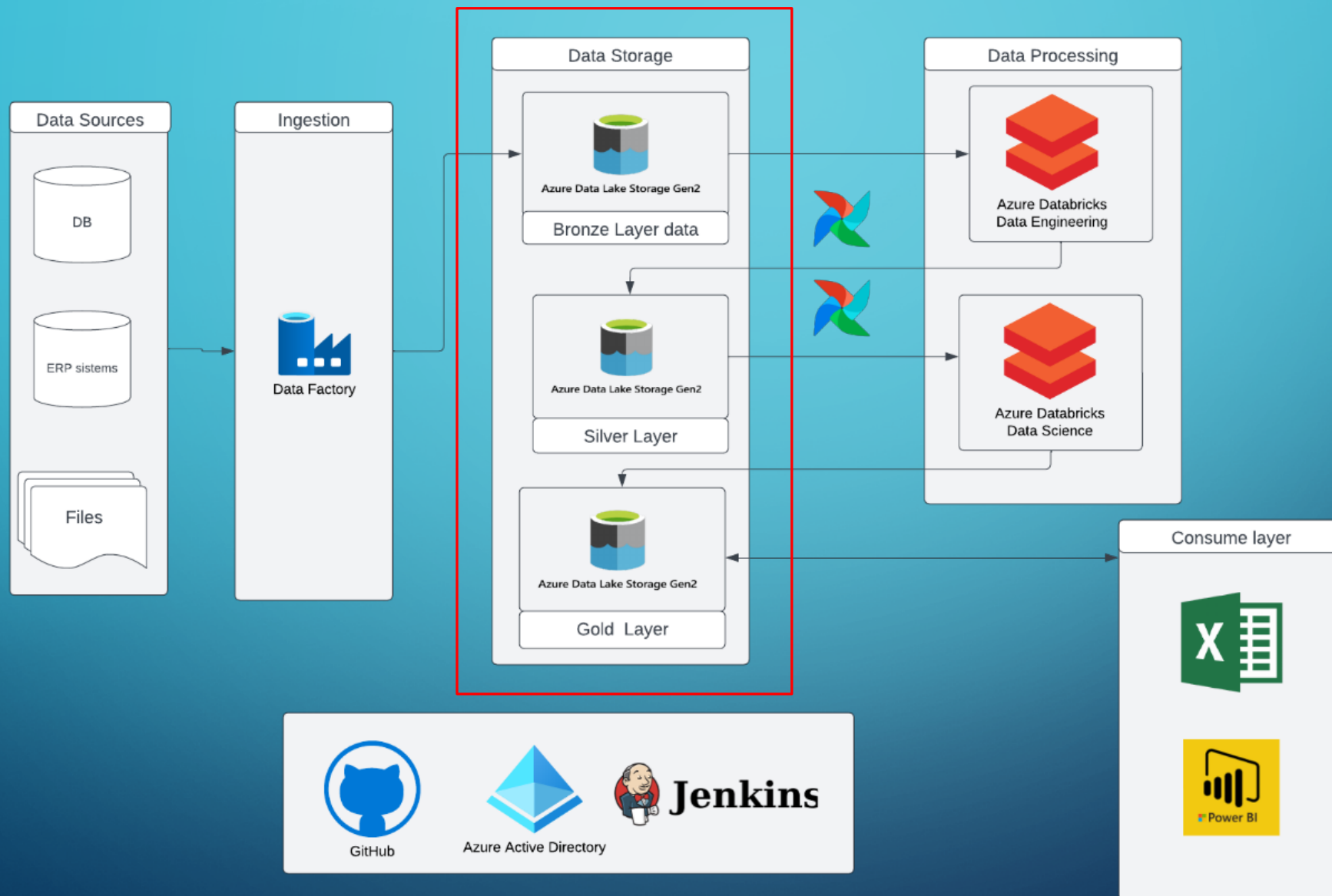
BASADA EN LA GENERALITZACIÓ, L'ESTUDI DE MERCAT, ELS OBJECTIUS D'ACTUALITAT, ESCALABILITAT, EFICIÈNCIA I SEGURETAT.

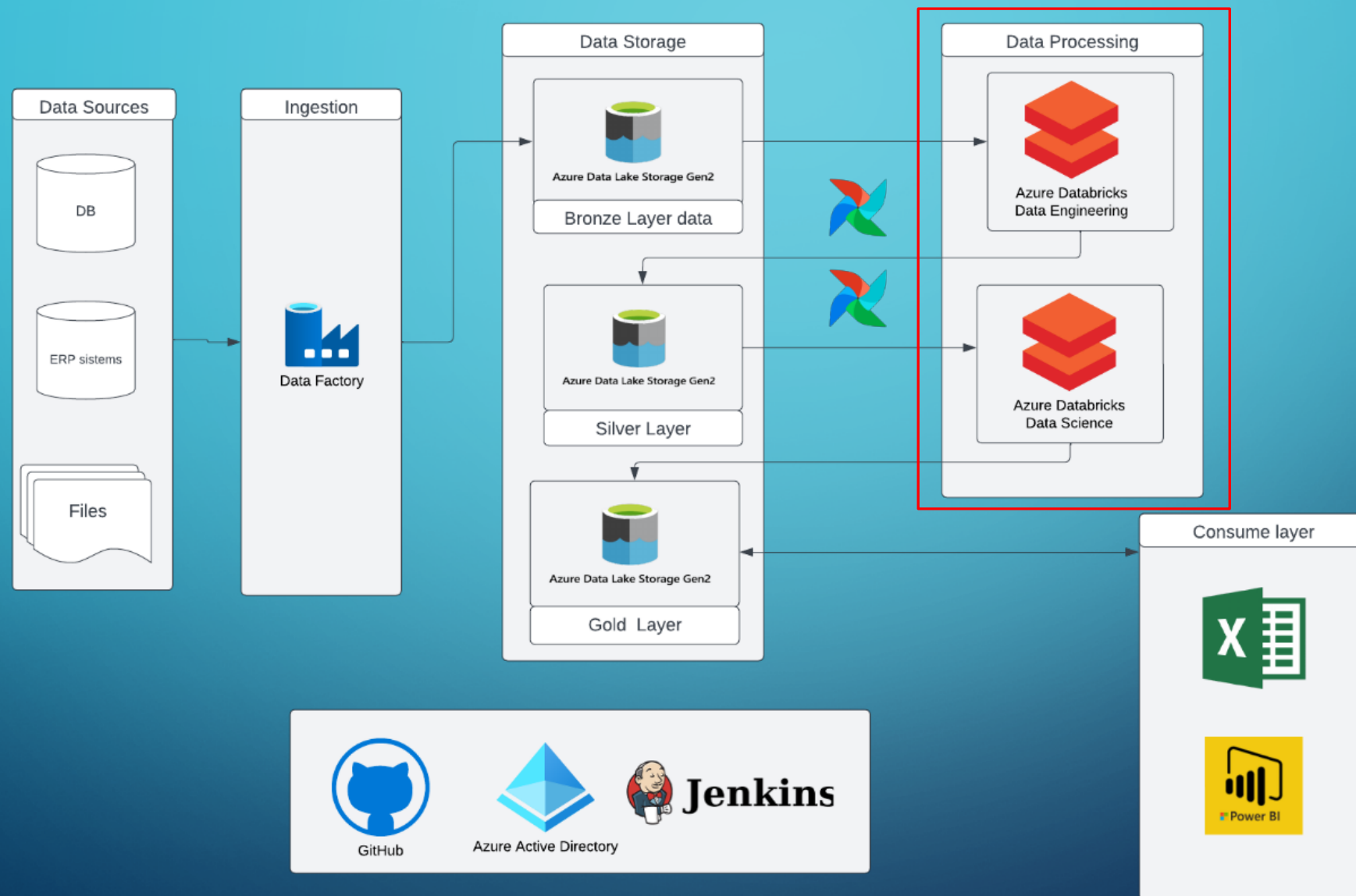
S'HA BUSCAT AÏLLAR LA SOLUCIÓ D'UN ÚNIC PROVEÏDOR.

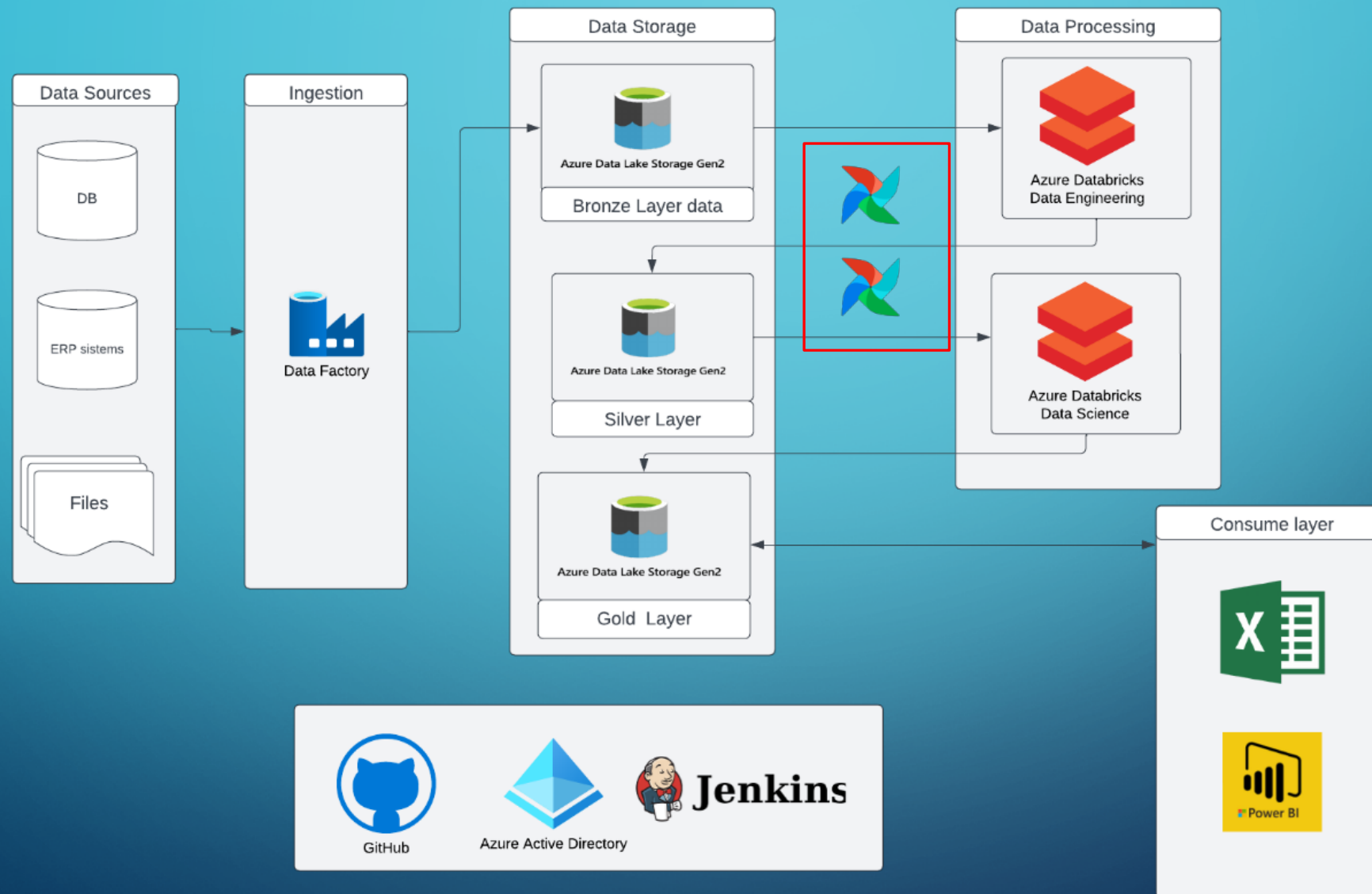


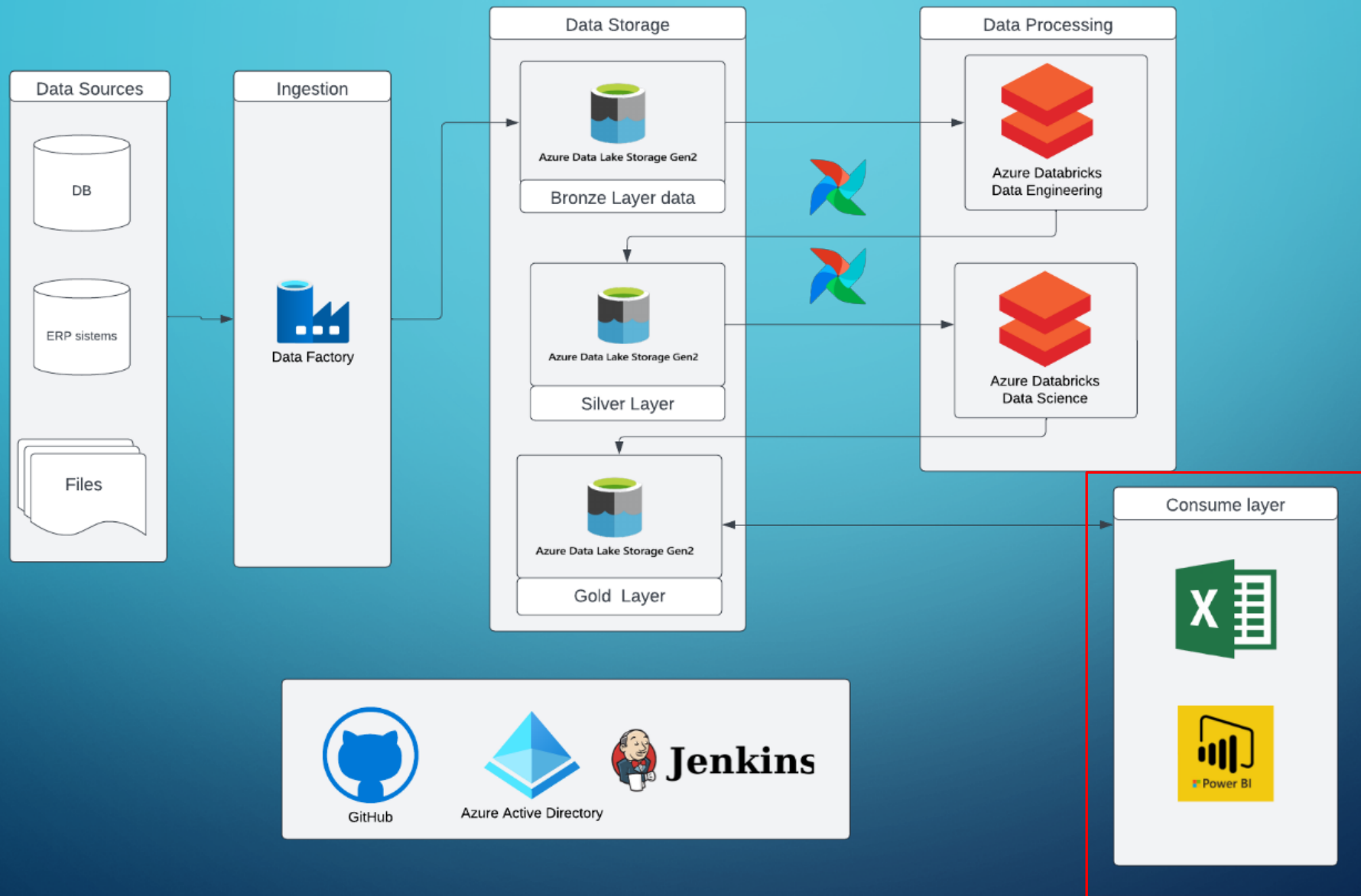


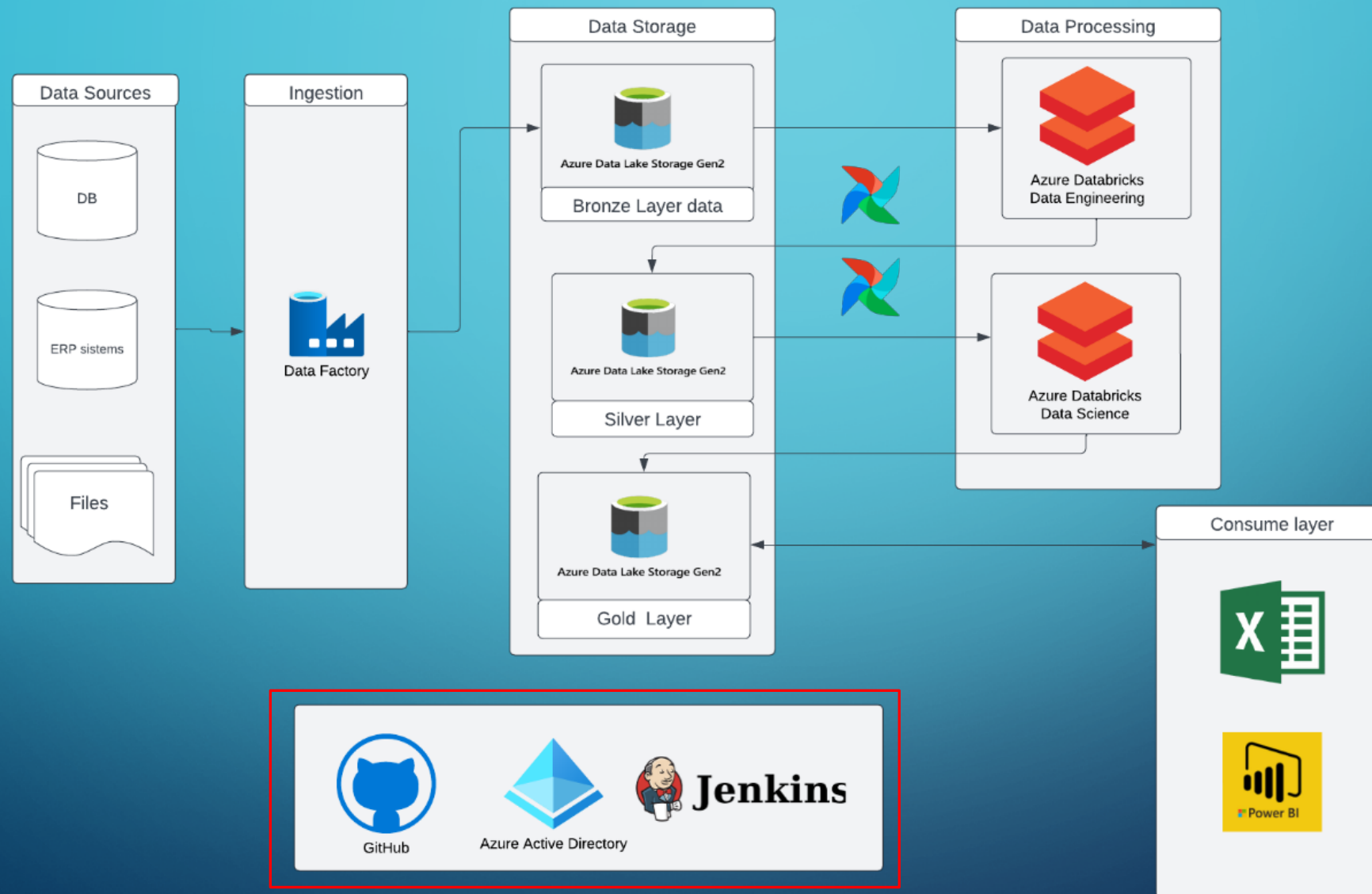












INTERCONEXIÓ I DESPLEGAMENT AUTOMÀTIC

- Control d'accés basat en Identitat (IAM) i Managed Identities a Azure.
- Connexió entre Airflow i Databricks mitjançant variables d'entorn.
- Docker-compose per a la orquestració i desplegament dels components.
- Connexió entre la aplicació de desplegament i Azure mitjançant Service Principal.

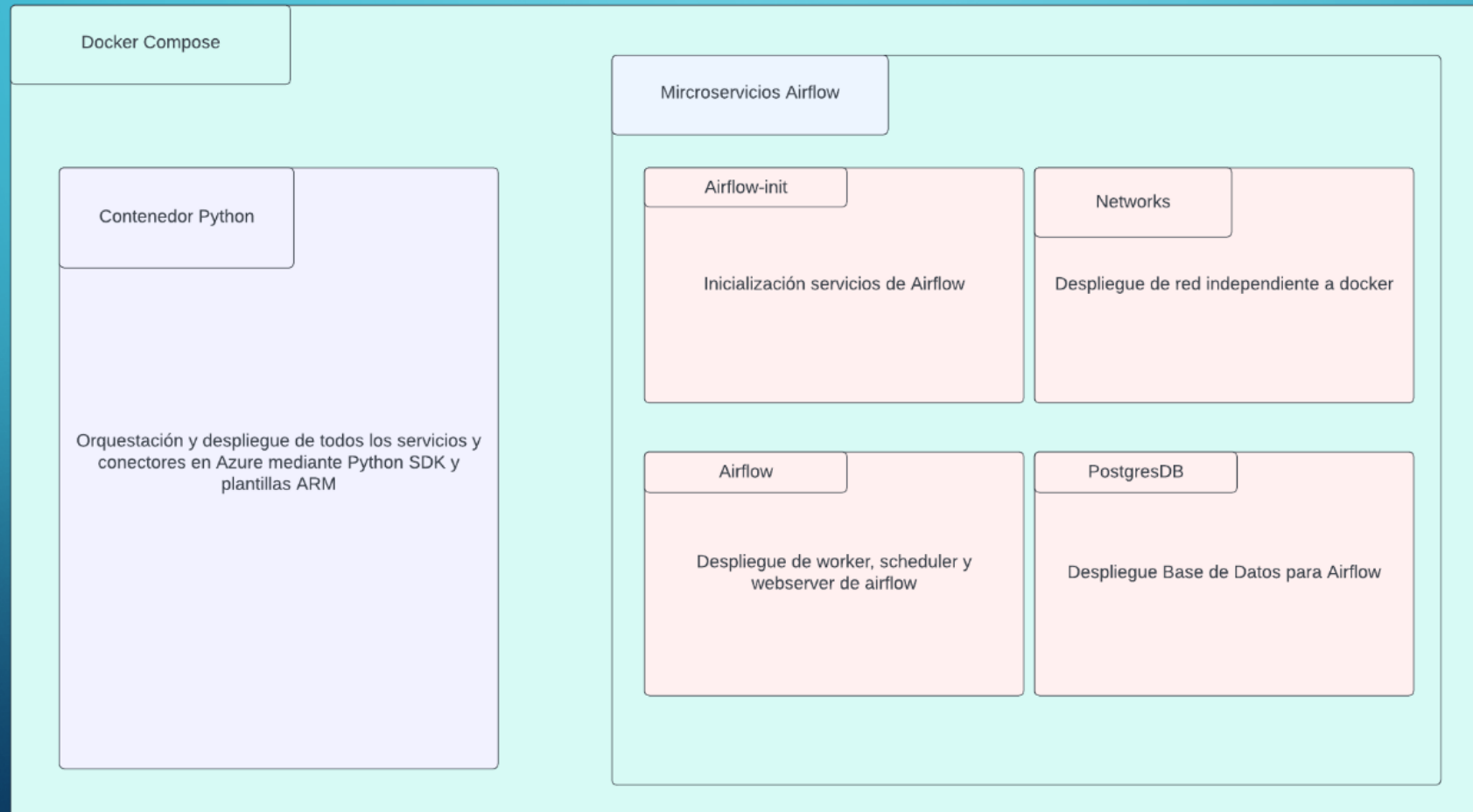
PRINCIPIS DE DISSENY DEL EXECUTABLE (IaC)

- L'script s'executa de manera atòmica per al desplegament de la infraestructura.
- Desplegament com a microserveis: Azure i Airflow.
- Ús de plantilles ARM i l'SDK d'Azure per Python per al desplegament de dels components al cloud.

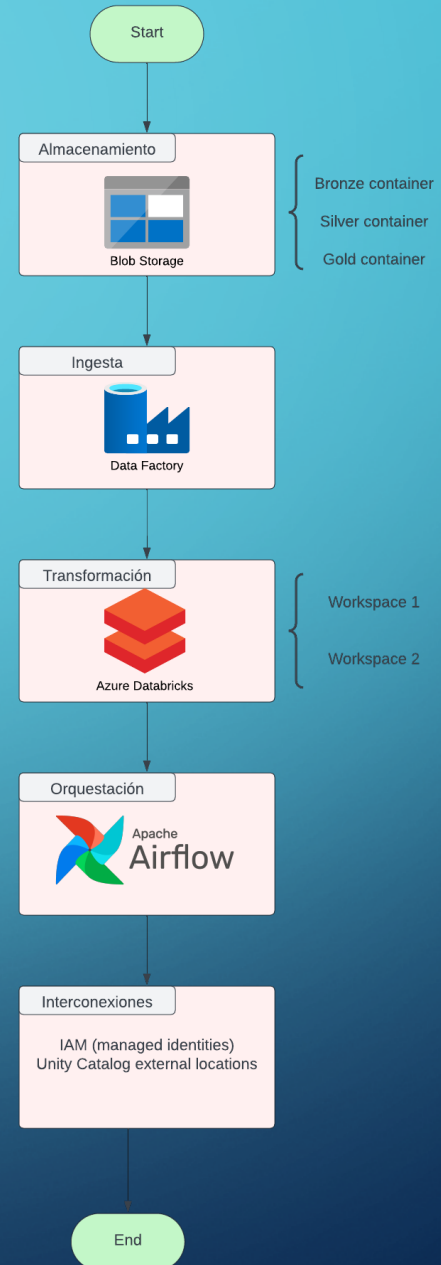
REQUERIMENTS

- Subscripció activa d'Azure.
- Instal·lació

FUNCIONAMENT DESPLEGAMENT DOCKER-COMPOSE



ORDRE DESPLEGAMENT



The background is a blue gradient. In the corners, there are white line-art illustrations of circuit boards or neural networks, with lines and small circles representing components.

CAS D'ESTUDI

VALIDACIÓ DEL PROCÉS END-TO-END

CONTEXT

Escenari: Una organització pretén analitzar els patrons de venda registrats en els seus sistemes OLTP.

Dataset: Dataset de Kaggle anomenat 'Amazon Sales Dataset' amb més de 1000 productes d'amazon, les transaccions de venda per usuaris i els reviews a cada producte.

Objectiu: Realitzar un procés end-to-end amb les dades, on s'ingesten al lakehouse, es transformen passant per les capes bronze, silver i gold i finalment es realitza un petit anàlisi de les dades netes per presentar en gràfiques.

INGESTIÓ DE DADES



<input type="checkbox"/>	adl-bronzelayer-test01	9/14/2023, 2:08:52 PM	Private	Available	...
<input type="checkbox"/>	adl-goldlayer-test-01	9/13/2023, 8:20:36 PM	Private	Available	...
<input type="checkbox"/>	adl-silverlayer-test-01	9/13/2023, 8:20:36 PM	Private	Available	...
<input type="checkbox"/>	source-test	9/14/2023, 5:56:59 PM	Private	Available	...

Data Factory interface showing the 'sales_copy' pipeline configuration. The 'Source' tab is selected, displaying a 'Copy data' activity with the source dataset 'DelimitedText1'.

Factory Resources:

- Pipelines: 1
 - sales_copy
- Change Data Capture (preview): 0
- Datasets: 2
- Data flows: 0
- Power Query: 0

Activity: Copy data

Source dataset: DelimitedText1

adl-bronzelayer-test01 container interface showing the 'test_table' and 'amazon_sales.parquet' files.

Authentication method: Access key (Switch to Azure AD User Account)

Location: adl-bronzelayer-test01

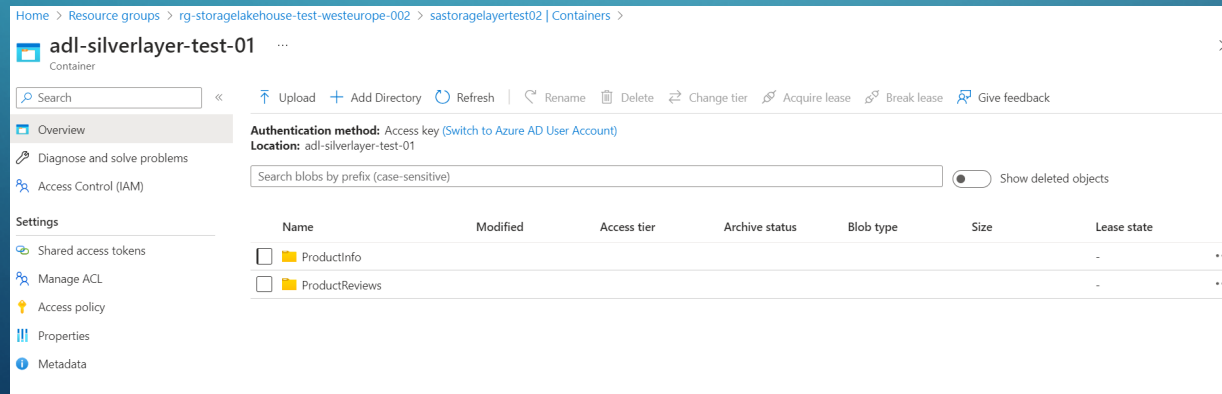
Search blobs by prefix (case-sensitive):

Name	Modified	Access tier	Archive stat
test_table			
amazon_sales.parquet	9/14/2023, 7:07:29 PM	Hot (Inferred)	

PREPROCESSAT DE DADES (BRONZE A SILVER)

Mitjançant un notebook de Databricks es realitzen les següents transformacions:

- Actualització tipus de dades
- Particionament
- Definició estructura de taules / fitxers
- Canvi format a Delta

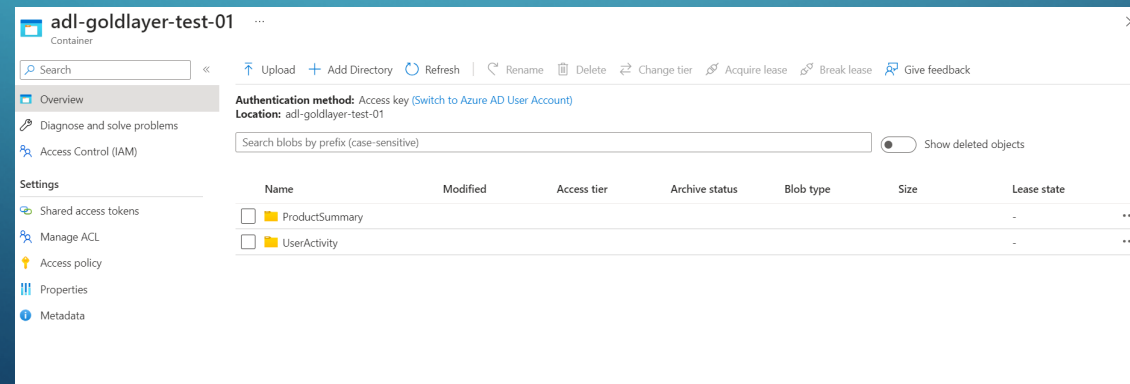


TRANSFORMACIÓ I ENRIQUIMENT DE DADES

(SILVER A GOLD)

Mitjançant un notebook de Databricks es realitzen les següents transformacions:

- Creació d'una nova taula **ProductSummary** agregada a partir de Joins de les taules Producte i Reviews on es calculen els rating mitjos per producte, el número total de reviews i el número de ventes.
- Creació d'una taula **UserActivity** on es guarden les activitats de compra i puntuació de cada usuari agregades.

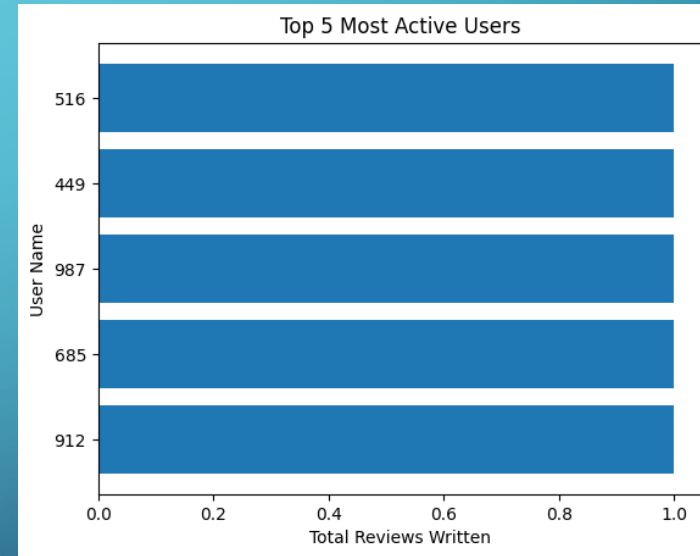


ANÀLISI DE DADES

Top 5 usuaris que donen més bons ratings

user_name ▲	total_reviews_written ▲	average_rating_given ▲
912	1	45
685	1	42
987	1	42
449	1	47
516	1	43

Top 5 usuaris que han fet més reviews



An abstract graphic on the left side of the slide, consisting of a network of light blue lines and small circles, resembling a circuit board or a neural network, set against a dark blue background.

CONCLUSIONS

- **Innovació:** Aquest projecte és un dels primers a abordar el tipus de flux de dades i arquitectura triats, la qual cosa el converteix en una font de referència per a futurs desenvolupadors en aquest àmbit.
- **Complexitat:** La manca de documentació i exemples previs ha fet que aquest projecte hagi suposat un repte tècnic gran. Per exemple, la connexió utilitzada per connectar l'emmagatzematge d'Azure i Databricks, que només fa uns pocs mesos que existeix.
- **Coneixements:** En el context del màster, aquest projecte té una cobertura integral del cicle de vida de les dades, oferint una comprensió més profunda del Big Data.
- **Resultat:** El cas d'estudi complet demostra la viabilitat de l'arquitectura triada i il·lustra el seu potencial per resoldre problemes reals.



An abstract graphic on the left side of the slide, consisting of a network of light blue lines and small circles, resembling a circuit board or a neural network, set against a dark blue background.

LIMITACIONES

- **Intervenció Manual Requerida:** Malgrat s'ha aconseguit una alta automatització, encara es requereix una certa intervenció humana en el procés, la qual cosa podria afectar l'eficiència i l'escalabilitat del sistema.
- **Barrera d'Entrada Tècnica:** La implementació del projecte requereix un coneixement tècnic significatiu, des de l'ús de contenidors Docker fins a la interacció amb Azure. Això limita l'accessibilitat per a aquells que manquen d'aquestes habilitats especialitzades.
- **Dependència d'Azure:** Al centrar-se en Azure com a proveïdor principal de serveis a la núvol, el projecte pot mancar de flexibilitat per integrar components d'altres proveïdors i limita les organitzacions a comprometre's amb l'ecosistema d'Azure.
- **Escalabilitat i Eficiència:** La necessitat d'intervenció manual a mesura que el sistema es expandeix podria conduir a ineficiències i errors humans, la qual cosa podria ser un repte en entorns empresarials on la velocitat i l'eficiència són crítiques.

An abstract graphic on the left side of the slide, consisting of a network of light blue lines and small circles, resembling a circuit board or a neural network, set against a dark blue background.

LÍNIES FUTURES

- 
- Execució en diversos proveïdors de cloud
 - Elecció de components personalitzats
 - Desplegament d'entorns de development i producció
 - Integració amb eines de CI/CD
 - Integració amb eines de control de versionat
- 

The background is a blue gradient. In the corners, there are white line-art illustrations of circuit boards or neural networks, with lines and small circles representing nodes.

PREGUNTAS