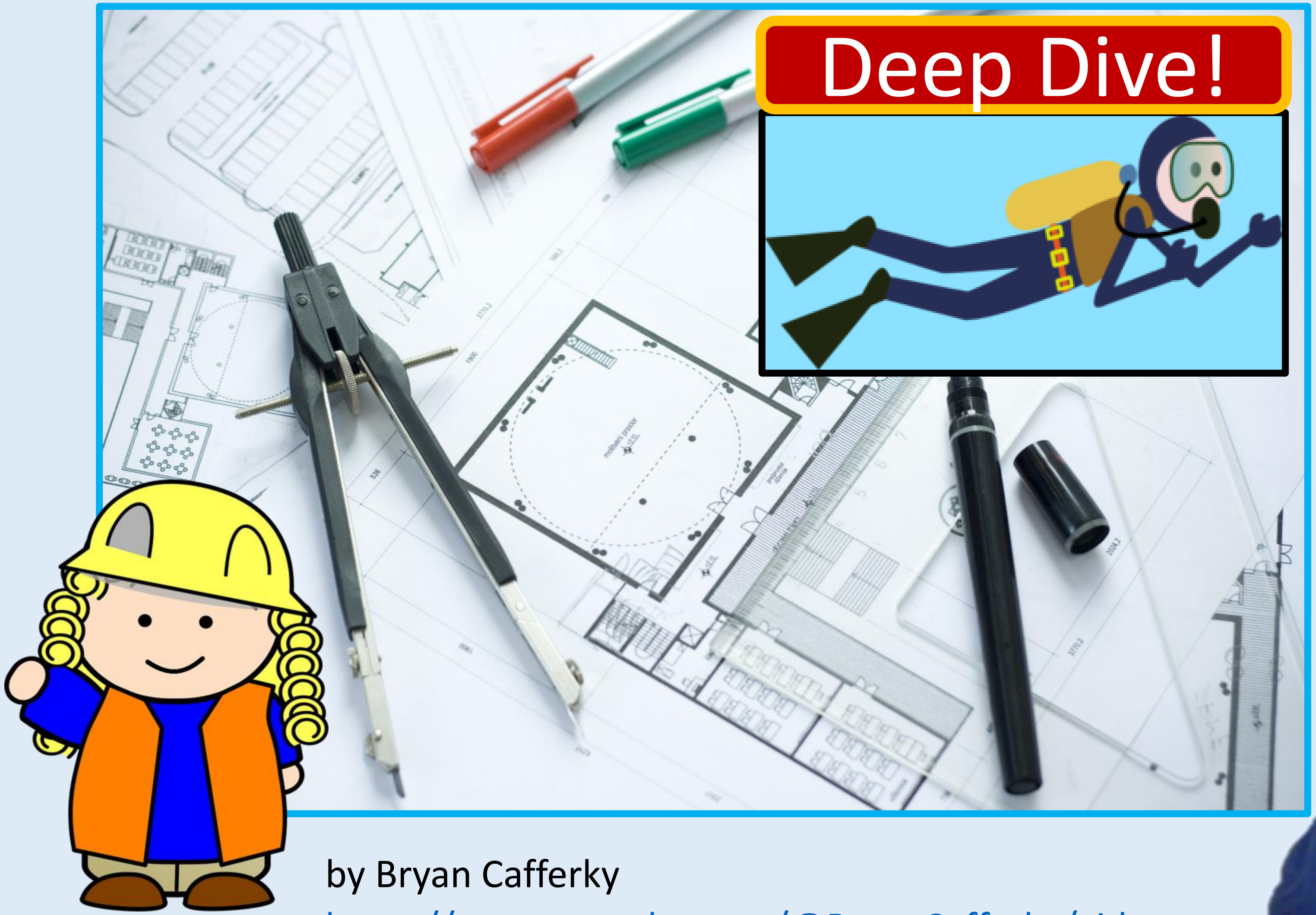
Data Architect vs. Data Engineer



by Bryan Cafferky
https://www.youtube.com/@BryanCafferky/videos

Where We're Going?

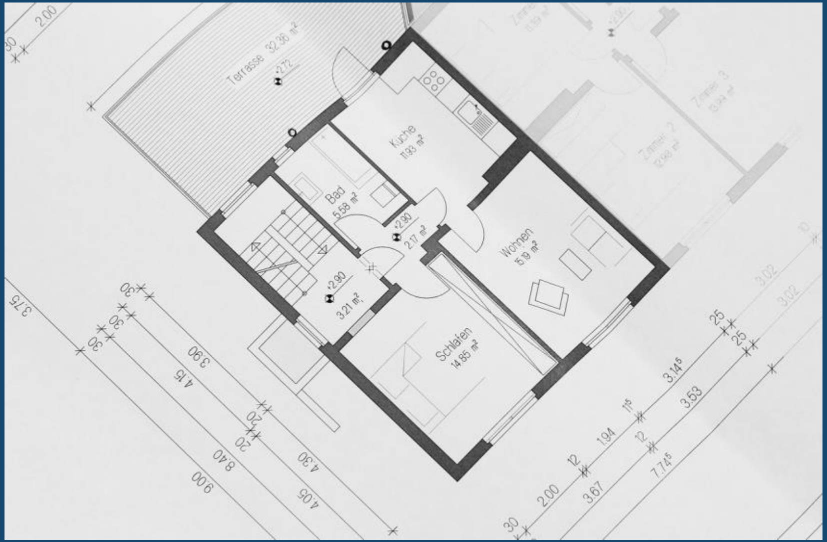
- Beware of Titles
- Reference Architecture vs. Solution Architecture
- What Does a Data Architect Do?
- What Does a Data Engineer Do?
- Avoiding Risks
- What About Data Scientists and Report Devs?
- Wrap Up

Beware of Job Titles

- **Architect in the Title Often is Inaccurate.**
- **Salespeople and Managers Often Have Architect Titles.**

Architect vs. Engineer

Data Architect



Data Engineer

Data Architect

- **Gather the Project Requirements.**

Where are you now? Where do you want to be?

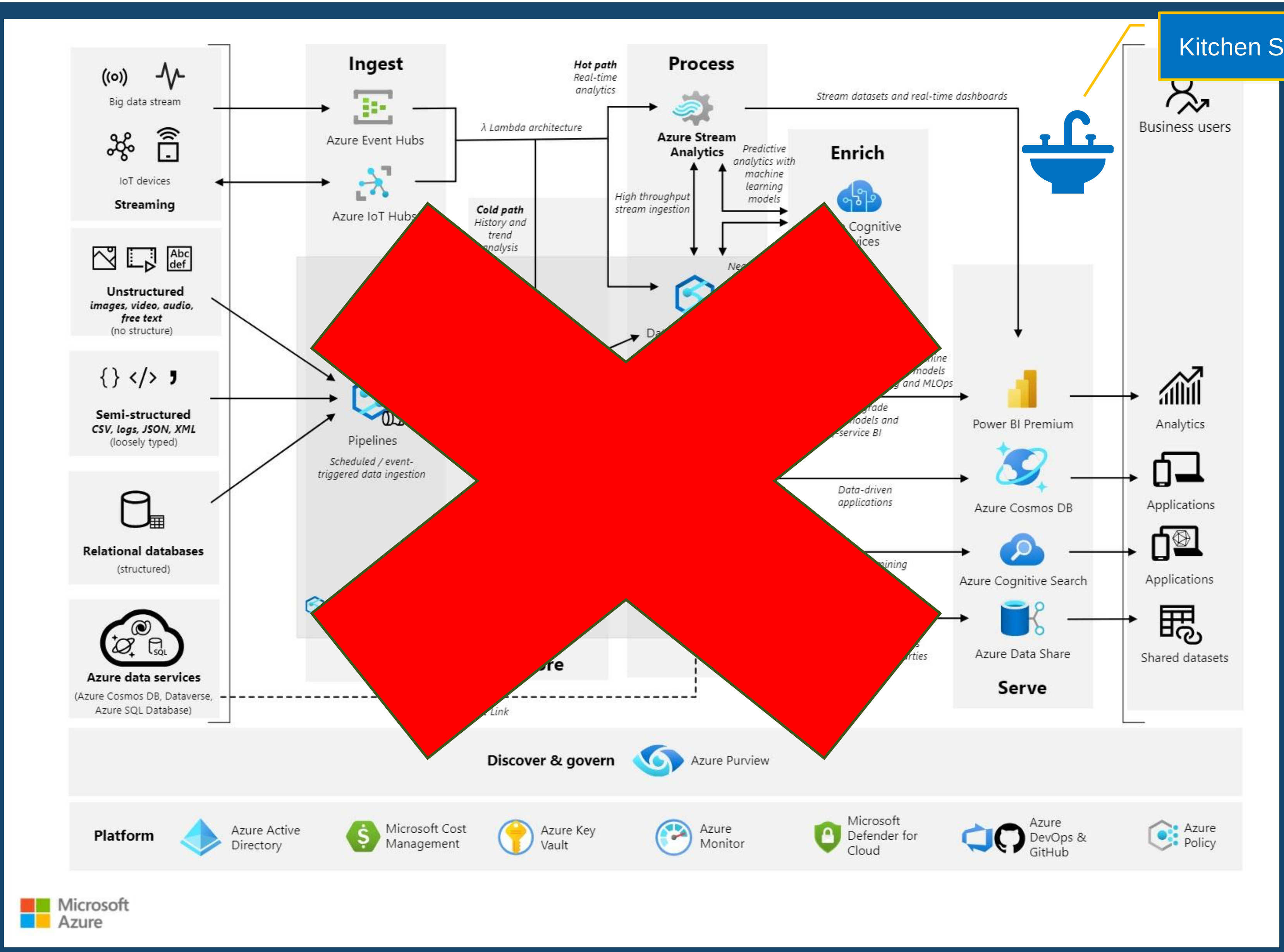
  - What is the business trying do? (AI, DW, Visualizations, Monetization)
  - Technical Environment (Cloud, Tools/Services, Languages)
  - Data: Volume, Structure/Unstructured, Velocity
  - Pain Points, Challenges, Risks, Constraints, Budget, Scope, Stakeholders
  - Sign Off

What does the Solution look like?

- **Define the High-Level Architecture. Review/Refine.**
  - Data Platform(s)
  - Supporting Services (ETL/ELT, Secrets, Storage)
  - Data Flow(s)
  - Security - Get Security Architecture Involved

- **Define the Detailed Architecture**
  - Orchestration, 3rd Party Services, Network Architecture, DataOps

Reference Architecture Diagram



Kitchen Sink

Only Includes Microsoft Services

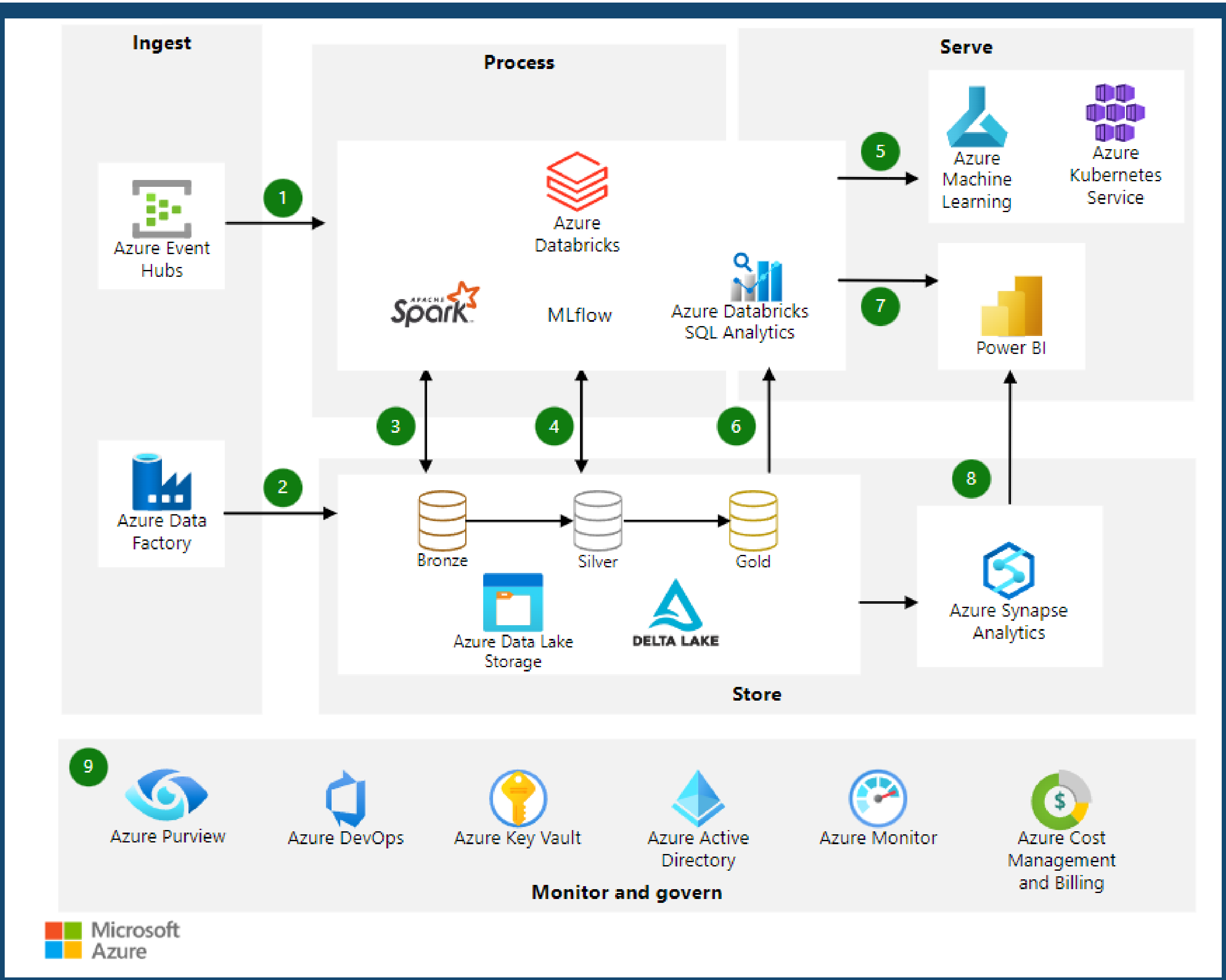
Throws Everything In

- Eye Chart.
- Designed to Sell.

Example Requirements

- ABC Investments wants some Power BI dashboards.
  - Want Customer Support reporting focused on identifying problems early so they can be resolved.
- Support data needs to be streamed in real time.
- Data Metrics:
  - 2 Million Events per Hour – support desk calls.
  - Some calls are critical issues.
  - Retain 6 Months of history.
  - Include Some Reference Data from Azure SQL Tables.
- Azure is the Preferred Platform.

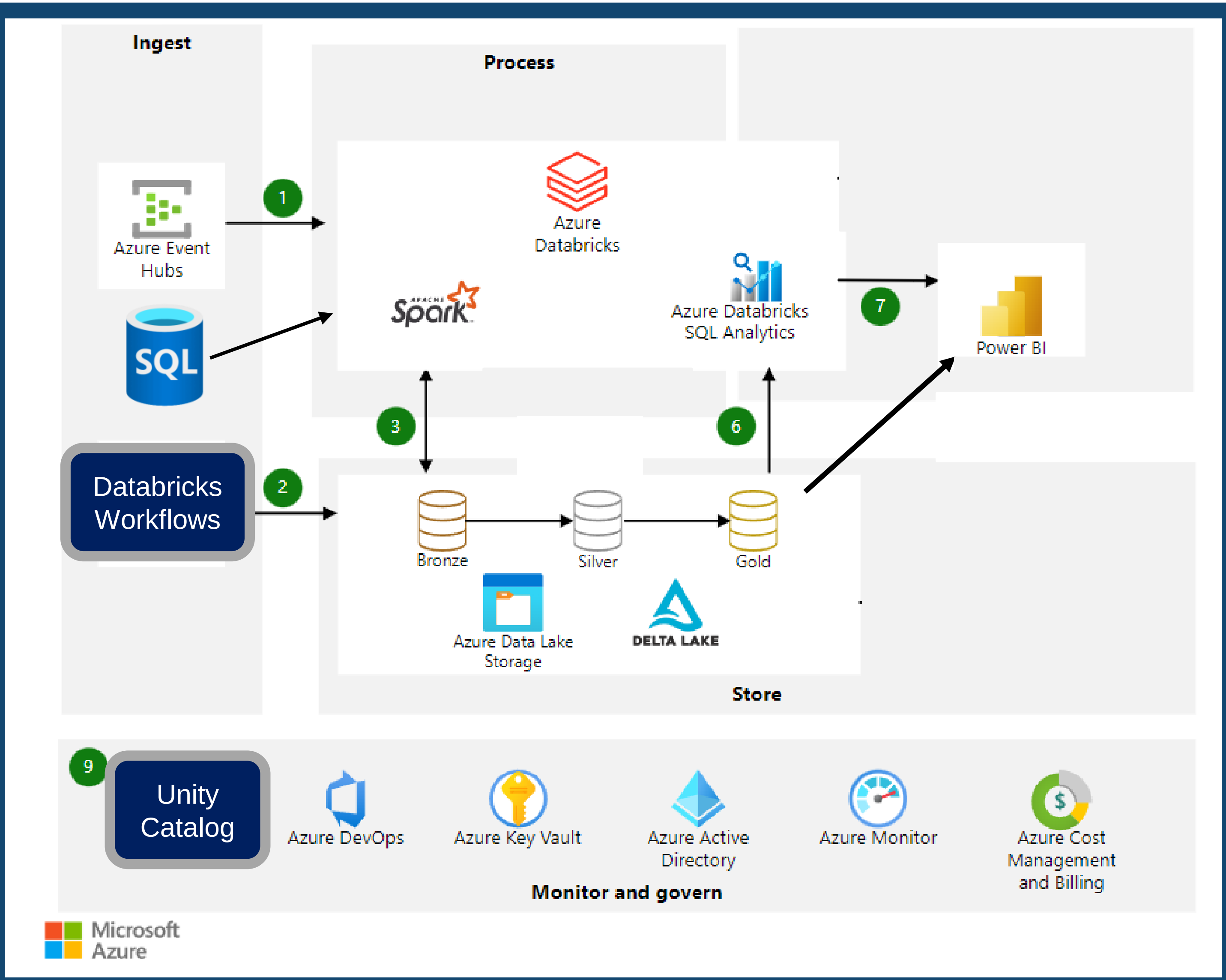
Solution Architecture Draft Example



Just Using This as a Starting Point.

https://learn.microsoft.com/en-us/azure/architecture/solution-ideas/articles/azure-databricks-modern-analytics-architecture

Solution Architecture Refined Example

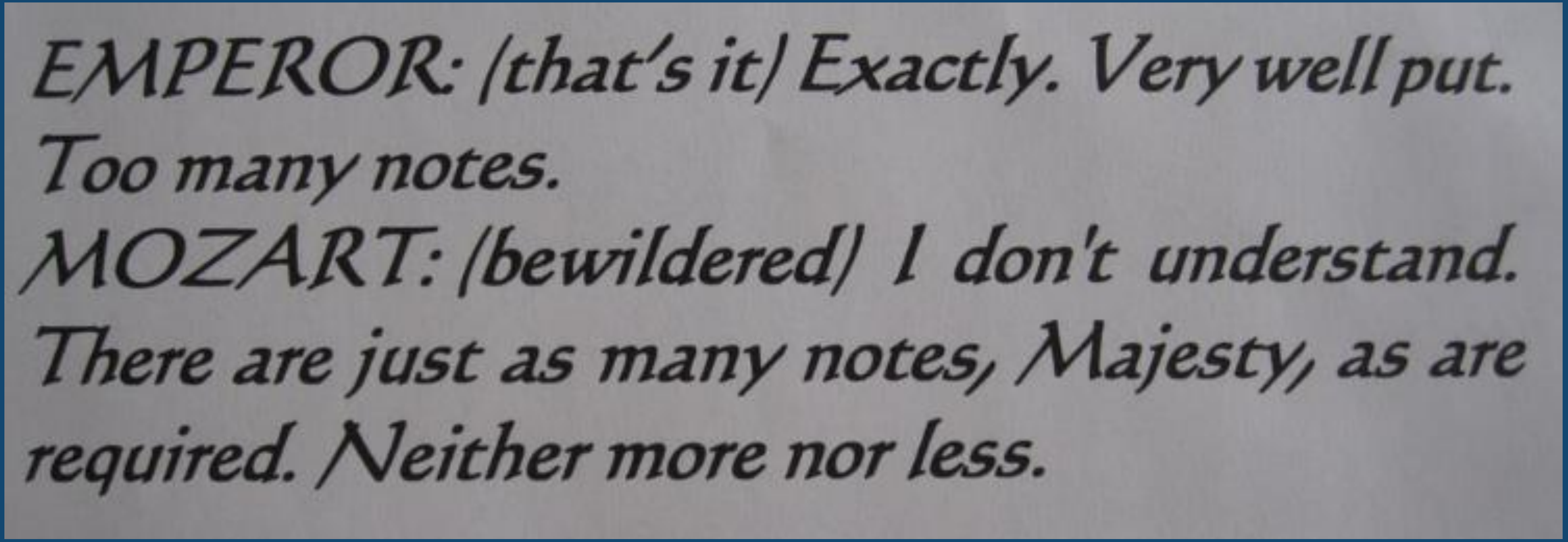


- ➢ Replaced ADF with Databricks Workflows.
- ➢ Replaced Purview with Unity Catalog.
- ➢ Removed Synapse (redundant).
- ➢ Removed Azure ML and Kubernetes.

✓ Simpler (KISS)
✓ Less Costly ($$$)
✓ Does Not Lose Any Required Functionality

https://learn.microsoft.com/en-us/azure/architecture/solution-ideas/articles/azure-databricks-modern-analytics-architecture

Parsimony & Mozart



EMPEROR: (that's it) Exactly. Very well put. Too many notes.

MOZART: (bewildered) I don't understand. There are just as many notes, Majesty, as are required. Neither more nor less.

Not too little and not too much. Lagom in Swedish.

Avoiding Unnecessary Risks

- Proof of Concepts
  - Develop a simple scaled down version of the architecture as a "sanity check" that it will meet the requirements. Fail early.
  - Test key requirements like row level security, frequent data refreshes, etc.

- Pilots & Minimum Viable Products (MVP)
  - Pick a subset of the solution functionality or business area to develop and deploy an initial phase of the solution. This limits the risks and helps identify problems earlier.
  - Pilots are REAL Deliverables!!!

Data Engineer

- Develop, Test, and Deploy Data Pipelines.

- Develop Pipelines.
  - Write Code to get data from sources, land it in storage (bronze), clean and transform it (silver), and aggregate it and save it to the solution layer (gold).

- Testing.
  - Run test data through the pipeline, and verify the output matches the requirements.

- Deploying Data Pipelines.
  - Automate Deployment via the appropriate tool (GitHub Actions, Azure DevOps, Scripts, Databricks Asset Bundles)

Which Is More Important?

Data Architecture

- Architecture is the Foundation Upon which You Build!
- Errors in Architecture Cost More to Fix!!!!
- The Earlier in the Process You Make Errors, the Costlier.
  Architecture -> Design -> Construction

What About?

Data Scientist

- Build ML Pipelines Off of the Data Pipelines.
- Restructuring the Data to Support Model Training.
- Construct ML Pipelines (train, evaluate, select, deploy) - MLOps

Report Developer

- Develop reports off the Data Pipelines.
- May Restructure Data to Support Reporting.
- Administration/Security Planning.

Wrapping Up

- Beware of Titles
- Reference Architecture vs. Solution Architecture
- What Does a Data Architect Do?
- What Does a Data Engineer Do?
- A [illegible]
- What About Data Scientists and Report Devs?

Thank You!