www.advancinganalytics.co.uk

SO YOU WANT TO BE A DATA ENGINEER?

@ADVANCINGANALYTICS @ADVANALYTICSUK /ADVANCING ANALYTICS

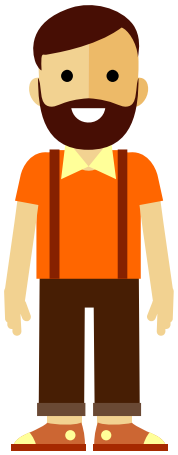
WHY ARE YOU HERE?


279900
99
INSERT 2 COINS
GUILE
ROUND 2
READY

BI Developer

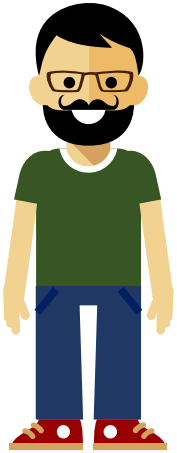
"Just a glorified ETL Developer"

"Just a developer that happens to focus on data"



Software Engineer

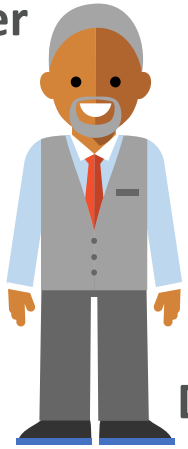
Data Engineer



Data Scientist

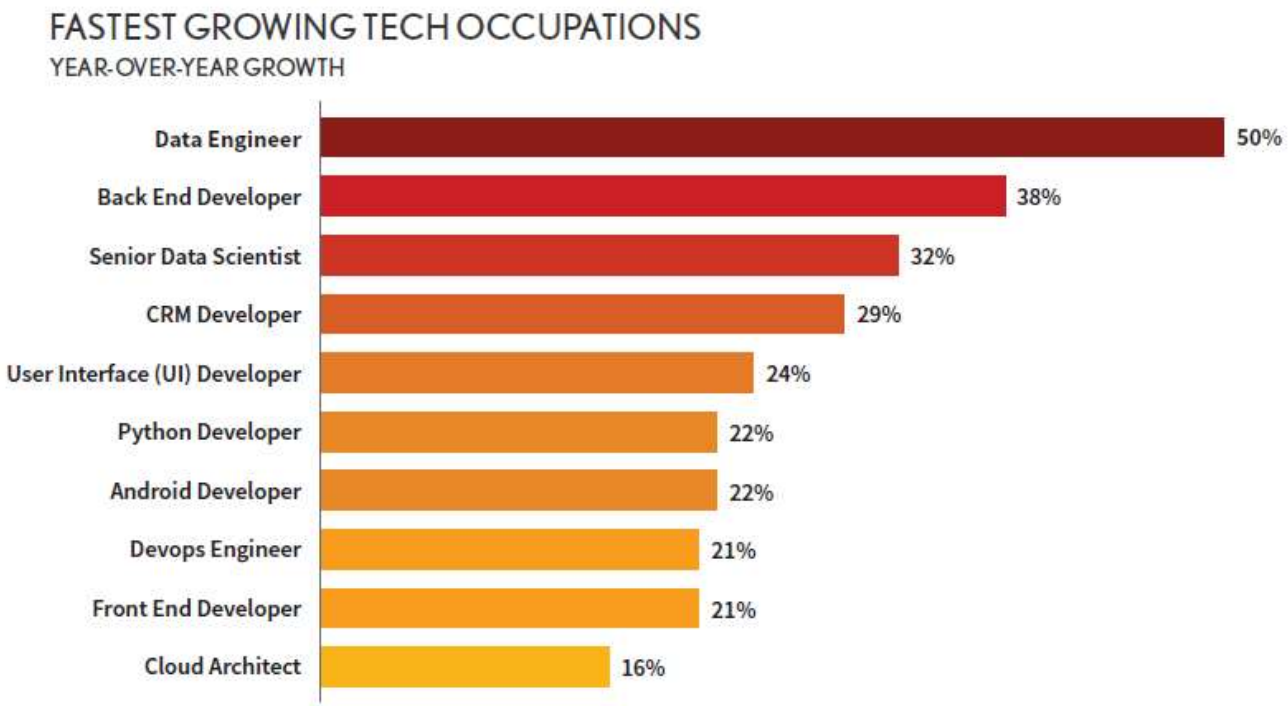
"A cog in the data science process"

"But we already built a Data Warehouse"



DBA/DRE

ADVANCING ANALYTICS


FASTEST GROWING TECH OCCUPATIONS
YEAR-OVER-YEAR GROWTH
Data Engineer
50%
Back End Developer
38%
Senior Data Scientist
32%
CRM Developer
29%
User Interface (UI) Developer
24%
Python Developer
22%
Android Developer
22%
Devops Engineer
21%
Front End Developer
21%
Cloud Architect
16%


The Software Engineer
The BI and Data Generalist
The DBA
POLICE OFFICER
LOS SANTOS POLICE
Est.1889
CHAOTIC GOOD

DATA ENGINEERING... IT'S FUN

ADVANCING ANALYTICS

HOW ARE WE GOING TO HELP

- WHAT IS DATA ENGINEERING
- THE TOOLS AND TECHNOLOGY
- How we Got Here

www.advancinganalytics.co.uk

WHAT IS A DATA ENGINEER?

BI Developer

- Warehousing
- Kimball
- Data Quality

Software Engineer

- Architecture Design
- DevOps
- Agile Development

Data Scientist

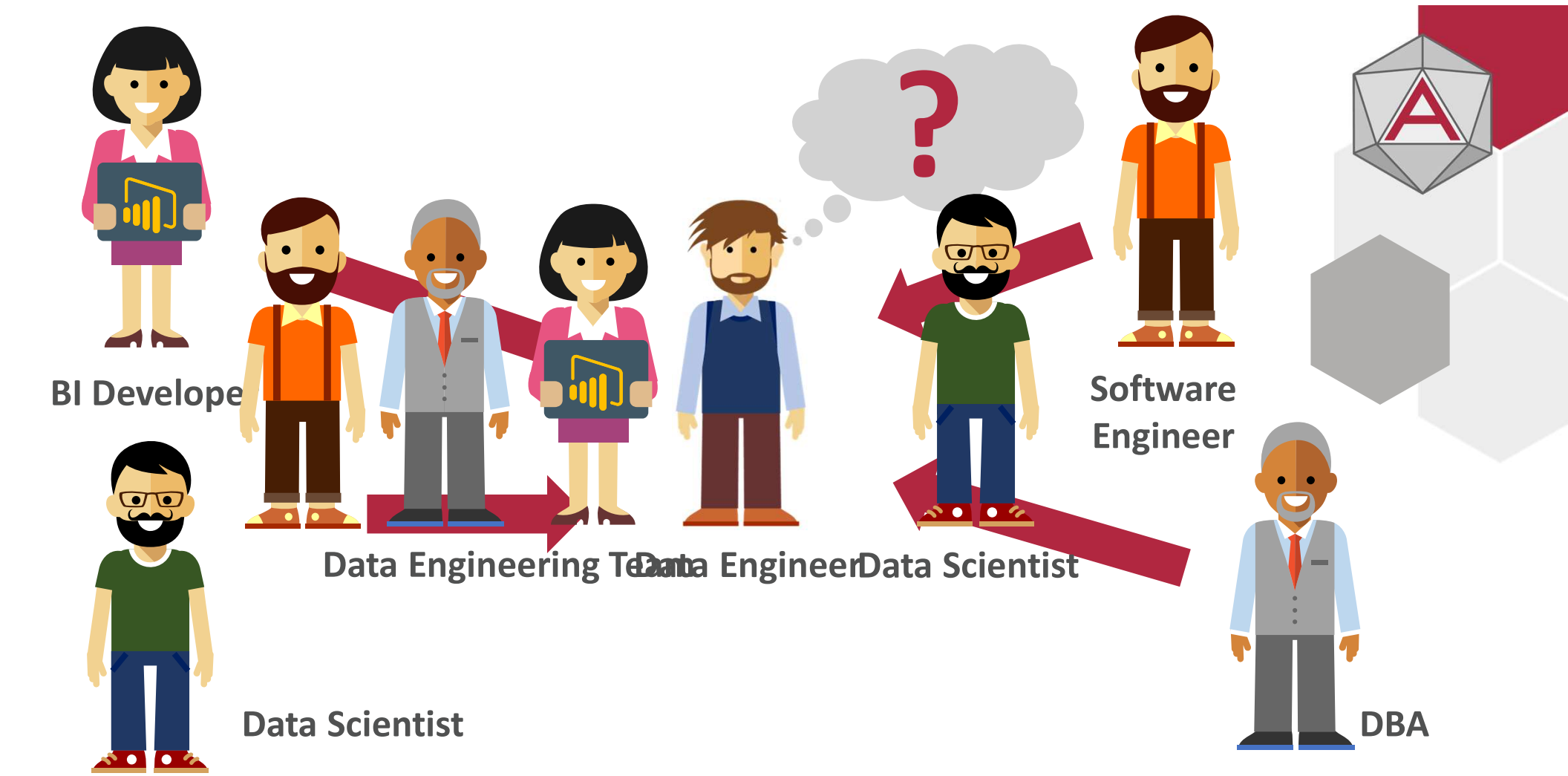
- Big Data Tools
- ML Engineering
- Exploratory Analysis

DBA

- Build Data Platforms
- SQL
- Performance & Security

Data Engineer

ADVANCING ANALYTICS


?
BI Develope
Data Engineering Team
Data Engineer
Data Scientist
Software Engineer
Data Scientist
DBA

WHAT DOES A DATA ENGINEER DO?

So is Data Engineering just about ETL (Extract Transform Load) Pipelines?

Talks to **stakeholders** and decides best **data architecture** for given project

Work with a variety of professionals to accomplish goals, including:

Data Scientists

Software Engineers

Business Analysts

ADVANCING ANALYTICS

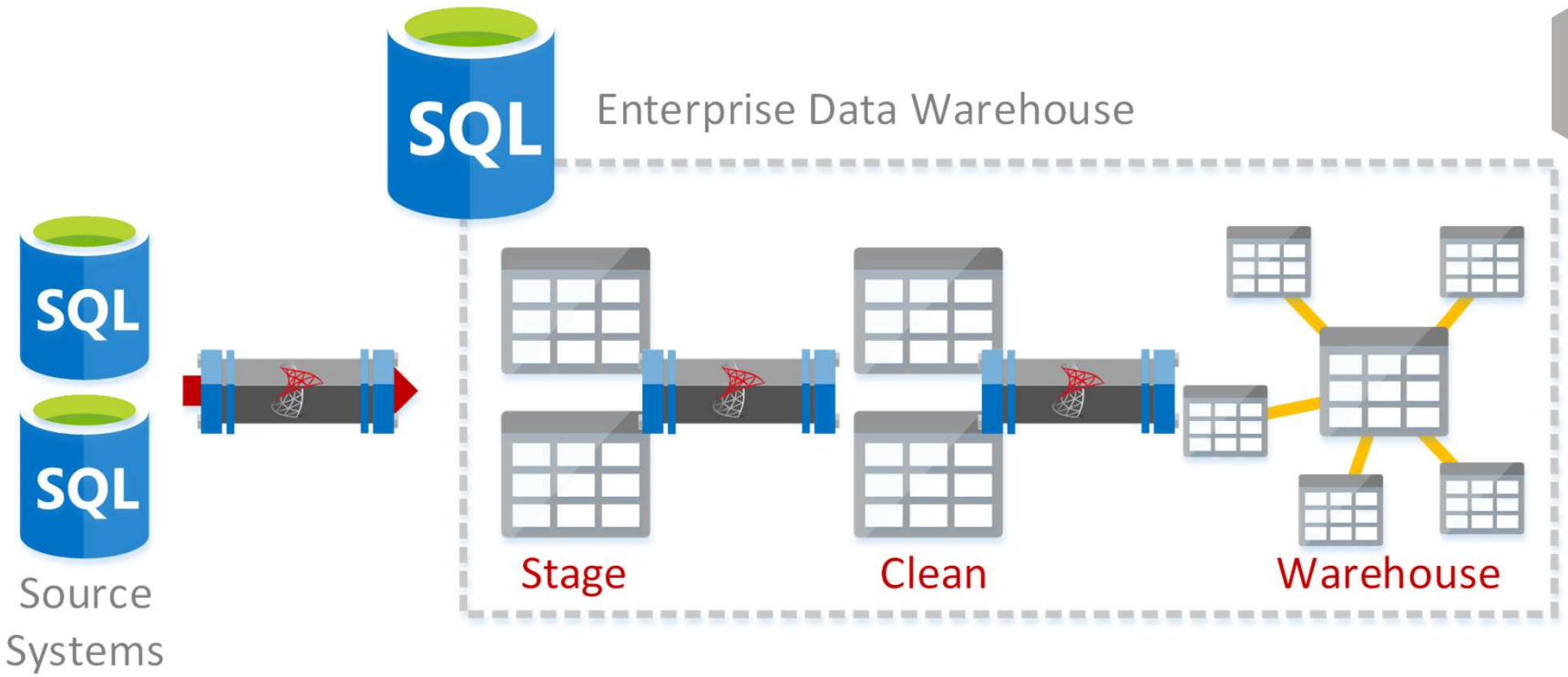
THE TOOLS AND TECHNOLOGY

WHAT IS DATA ENGINEERING

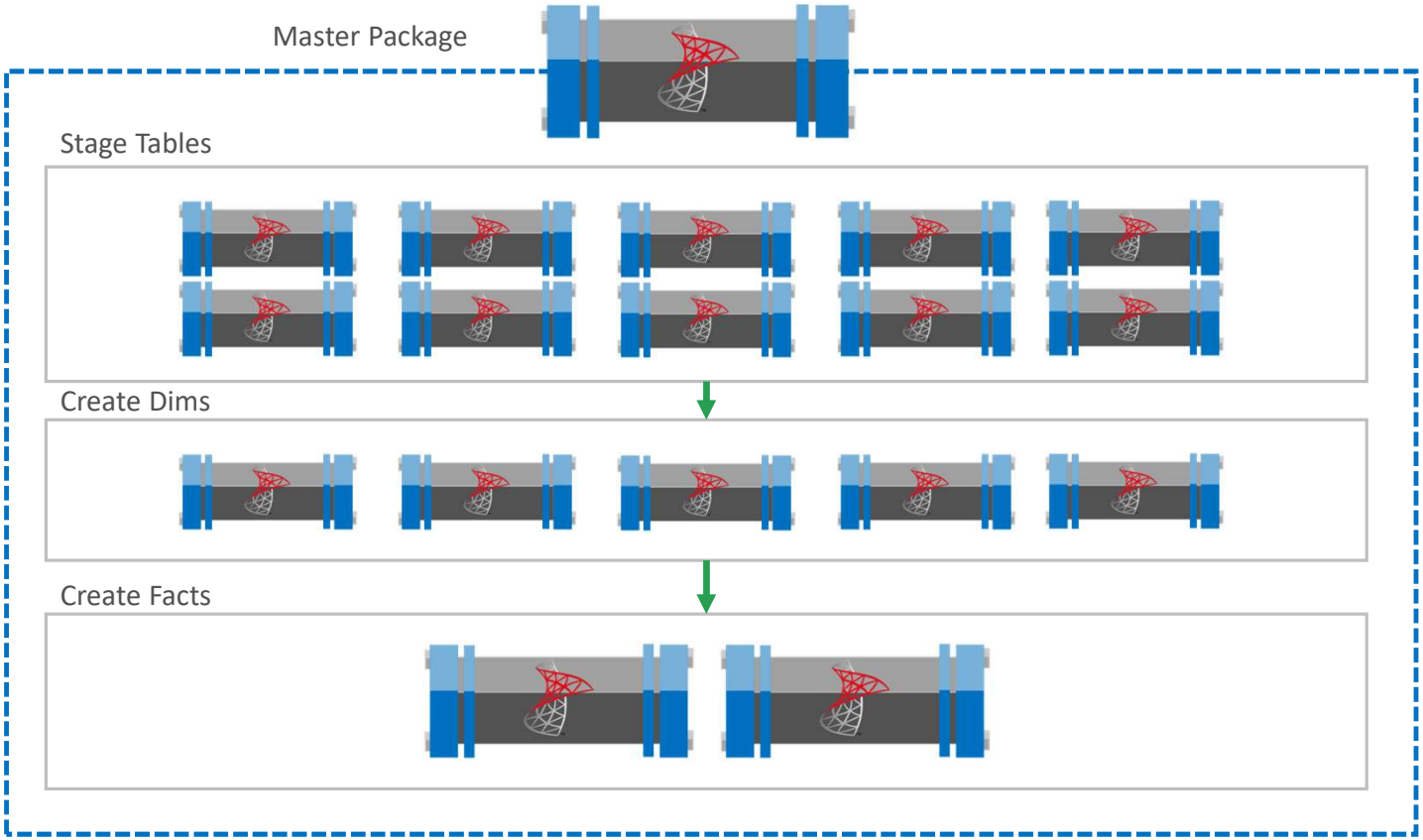
How we Got Here

WHERE IT ALL BEGAN

THE OLD WORLD



PACKAGE ORCHESTRATION



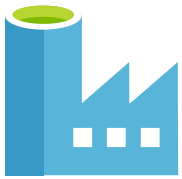
OUR ETL TOOL CHECKLIST

- Low Dev Effort
- Many Use Cases
- Flexible Formats
- Elastic Scaling
- Agile
- Support

FAILED

SO WE MOVE TO THE CLOUD RIGHT?

MORE REALISTICALLY, WE HAVE SEVERAL KEY OPTIONS



Azure Data Factory

Visual cloud-native Orchestration. Excellent at moving data between sources

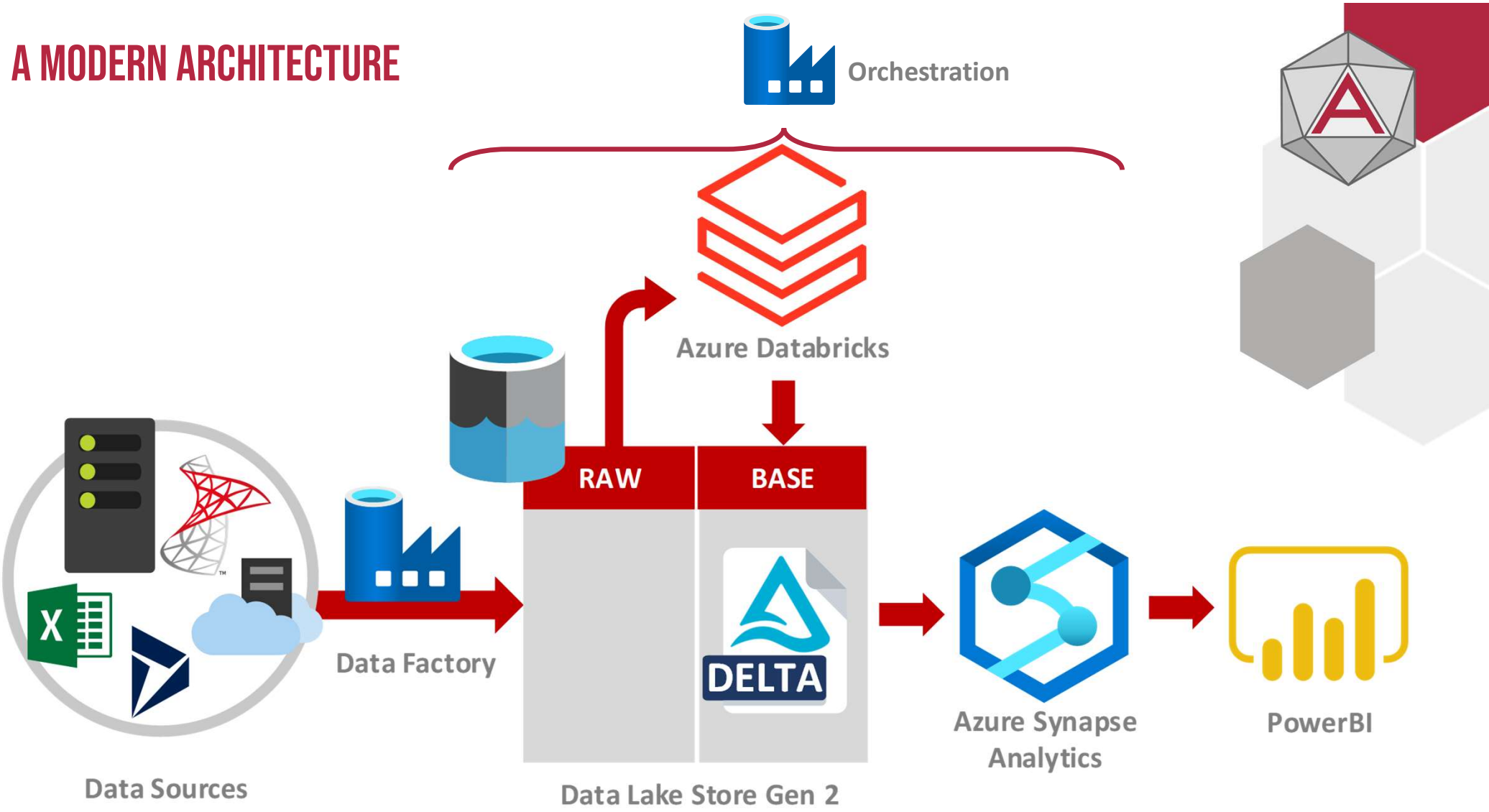
Databricks

Spark as a service, powerful multi-language big data engine. Supports SQL but best with Python/Scala

Azure Synapse Analytics

Suite of tools ranging from orchestration to engineering & full data warehousing

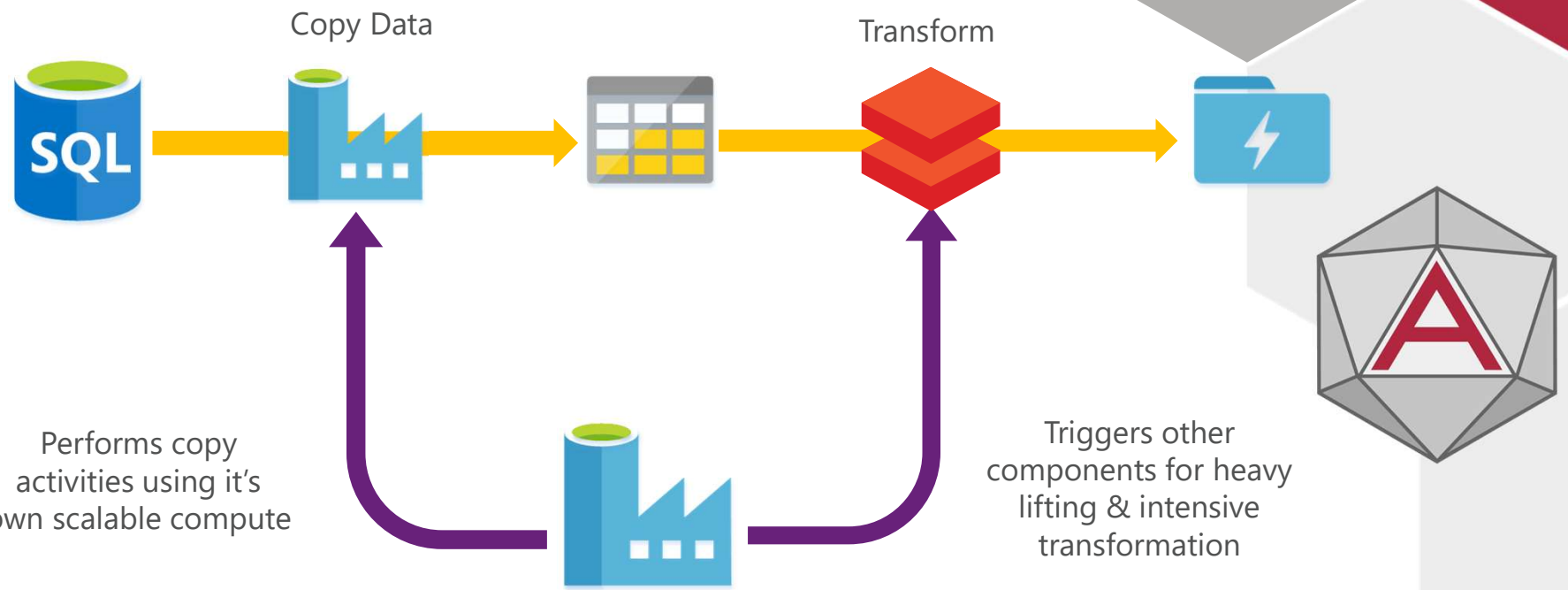
A MODERN ARCHITECTURE



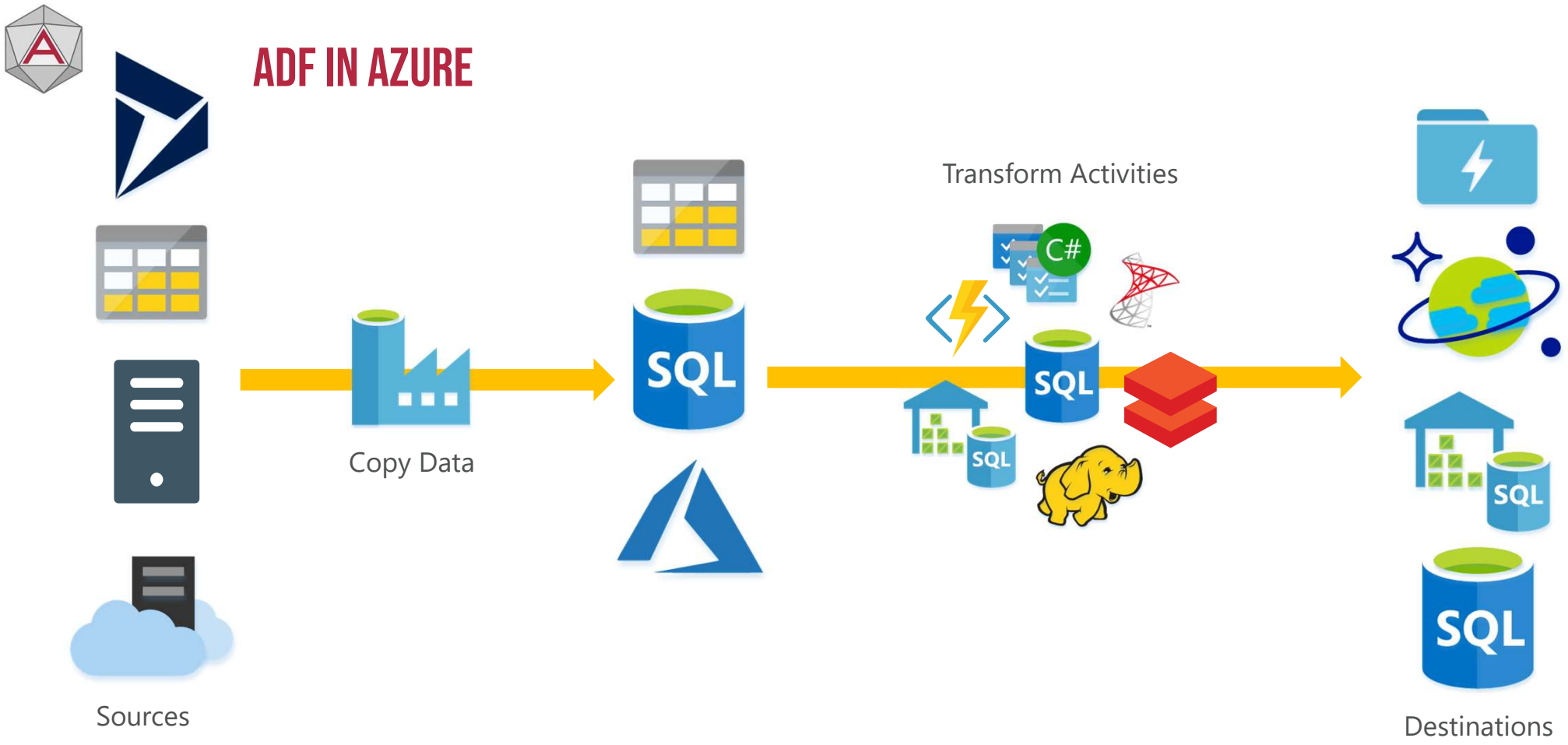
www.advancinganalytics.co.uk

ORCHESTRATION & INTEGRATION

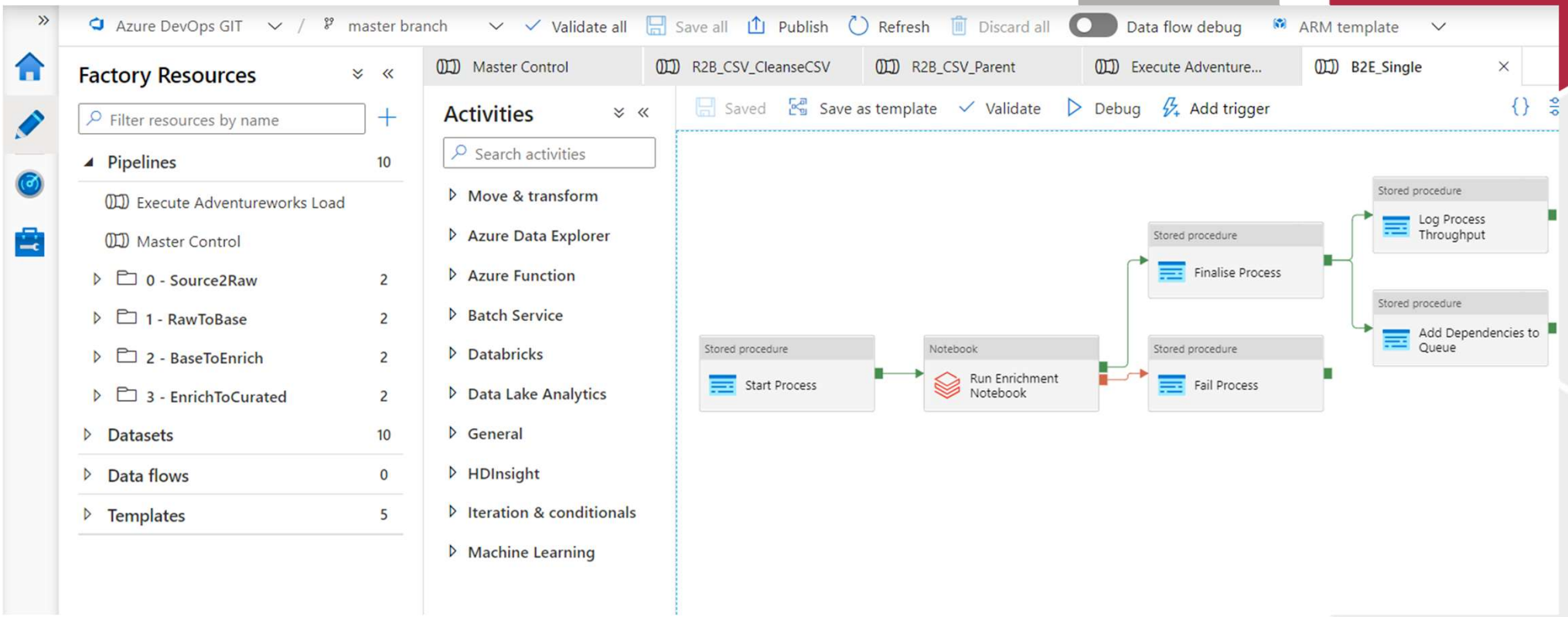
AN ADF PRIMER



ADF IN AZURE

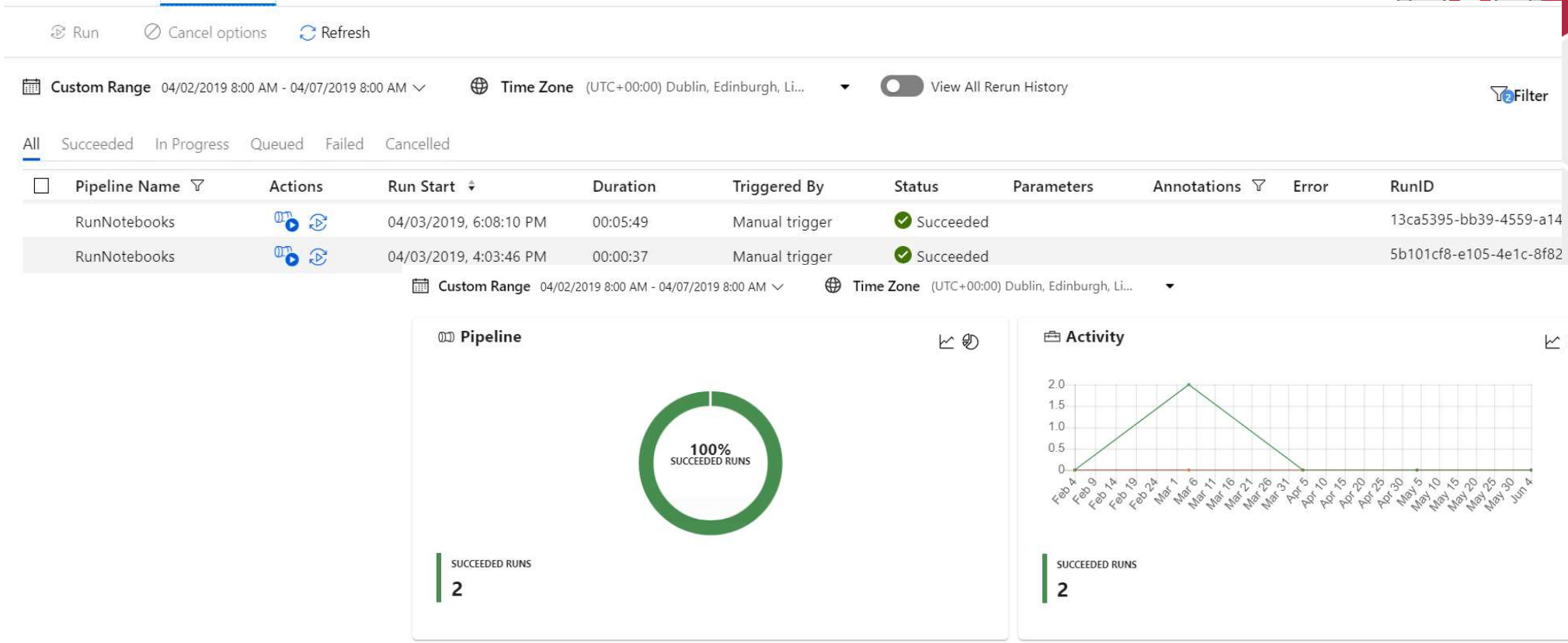

Transform Activities
C#
SQL
SQL
SQL
SQL
SQL
Copy Data
Sources
Destinations

DEVELOPER TOOLS

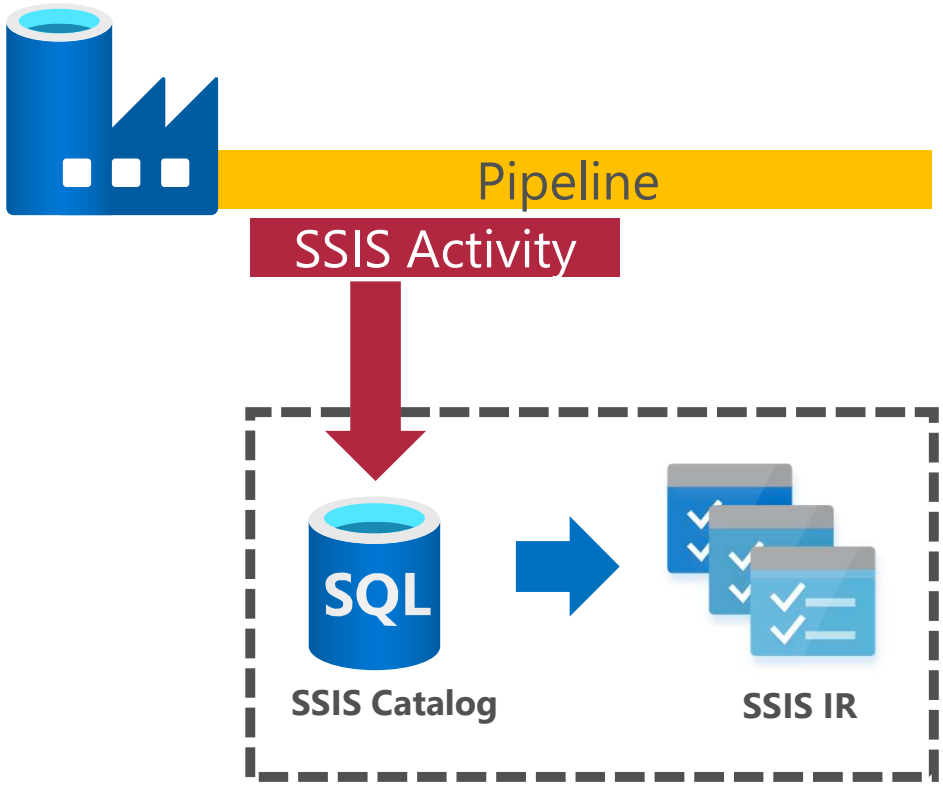


ADVANCING ANALYTICS

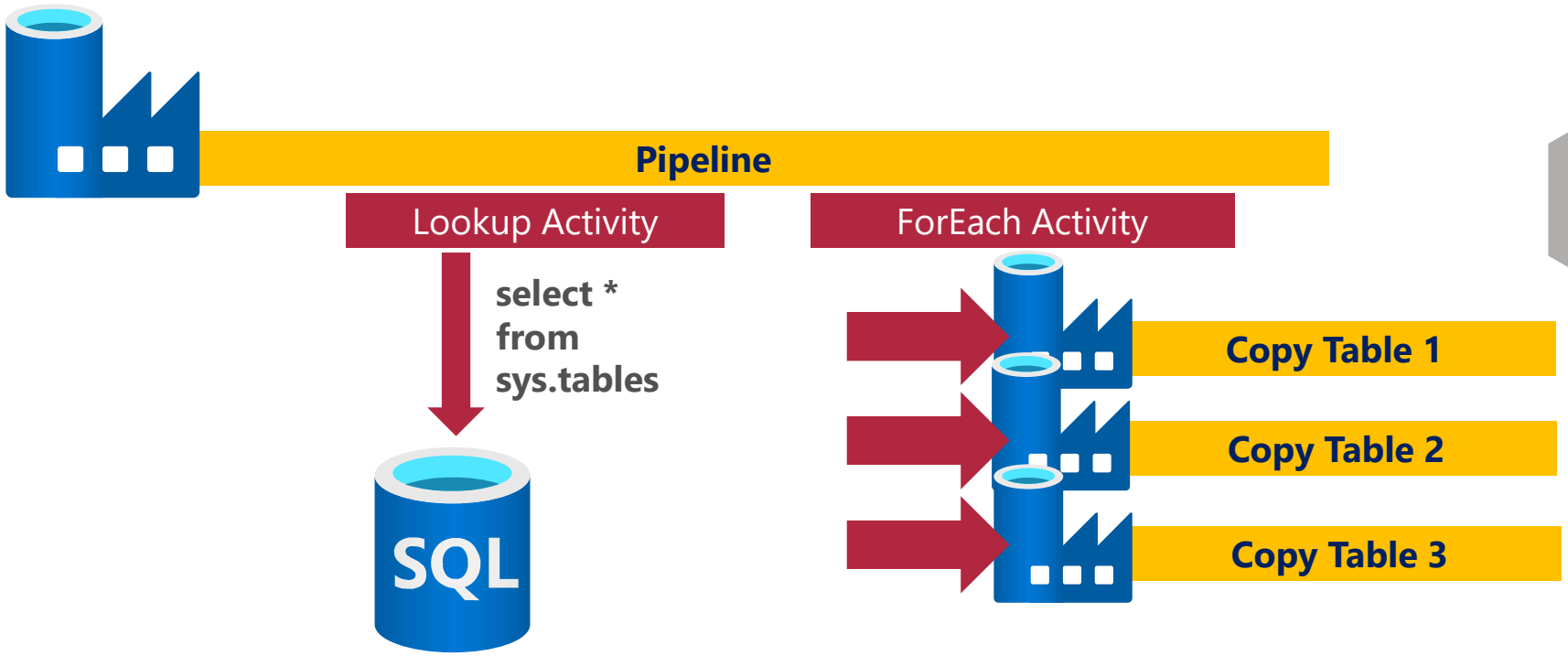
MONITORING


Run
Cancel options
Refresh
Custom Range 04/02/2019 8:00 AM - 04/07/2019 8:00 AM
Time Zone (UTC+00:00) Dublin, Edinburgh, Li...
View All Rerun History
Filter
All
Succeeded
In Progress
Queued
Failed
Cancelled
Pipeline Name
Actions
Run Start
Duration
Triggered By
Status
Parameters
Annotations
Error
RunID
RunNotebooks
04/03/2019, 6:08:10 PM
00:05:49
Manual trigger
Succeeded
13ca5395-bb39-4559-a14
RunNotebooks
04/03/2019, 4:03:46 PM
00:00:37
Manual trigger
Succeeded
5b101cf8-e105-4e1c-8f82
Custom Range 04/02/2019 8:00 AM - 04/07/2019 8:00 AM
Time Zone (UTC+00:00) Dublin, Edinburgh, Li...
Pipeline
100%
SUCCEEDED RUNS
SUCCEEDED RUNS
2
Activity
2.0
1.5
1.0
0.5
0
Feb 4
Feb 9
Feb 14
Feb 19
Feb 24
Mar 1
Mar 6
Mar 11
Mar 16
Mar 21
Mar 26
Mar 31
Apr 5
Apr 10
Apr 15
Apr 20
Apr 25
Apr 30
May 5
May 10
May 15
May 20
May 25
May 30
Jun 4
SUCCEEDED RUNS
2

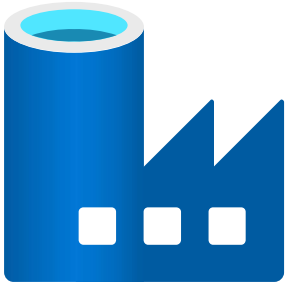
SSIS – LIFT & SHIFT POTENTIAL



USE THE FLEXIBILITY OF DATA FACTORY



DATA FACTORY RECAP



- Orchestrates all data workflows in Azure
- Best method of onboarding data to Azure
- Use parameters, forEach and child executions

www.advancinganalytics.co.uk

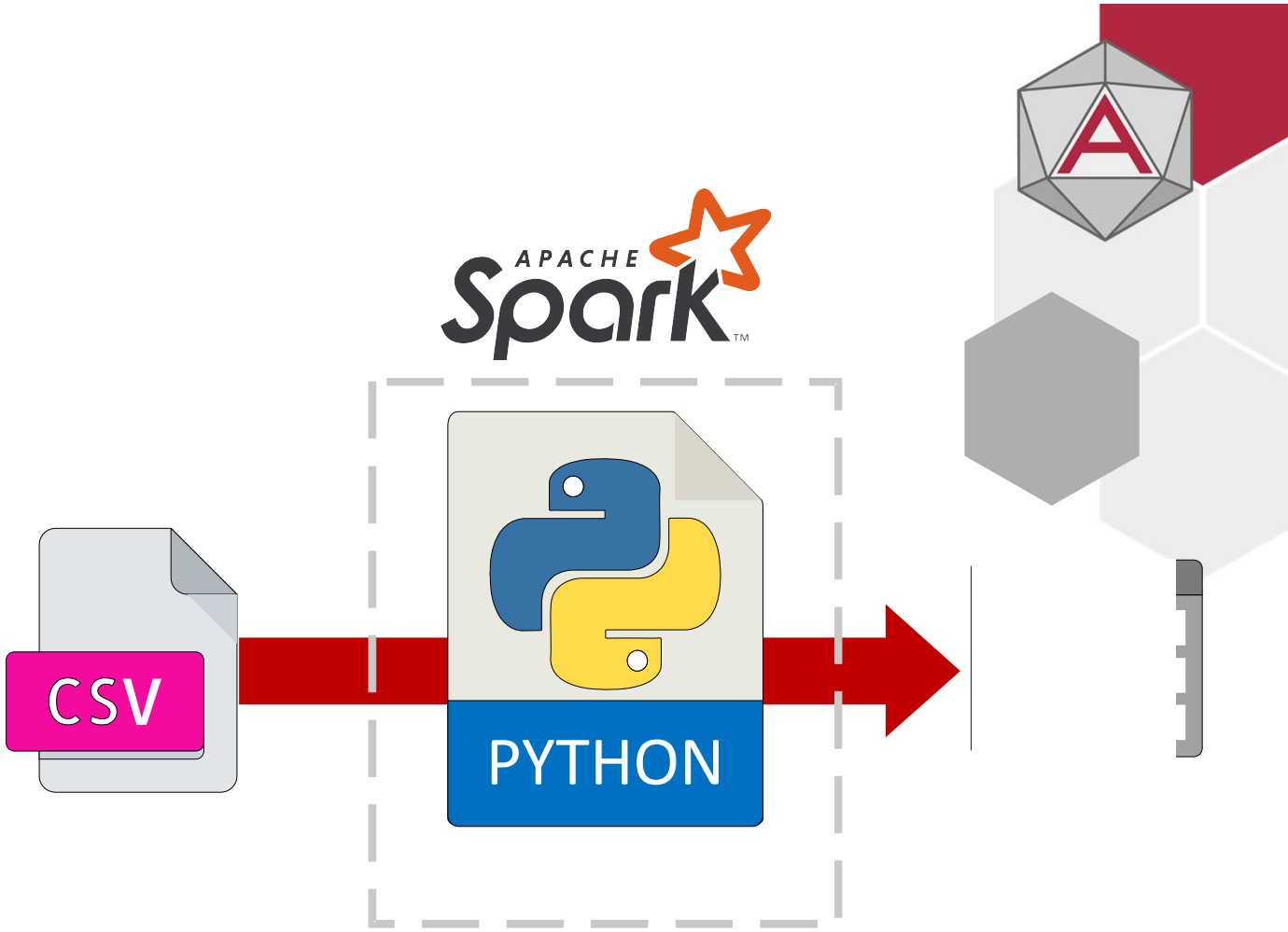
DATA PROCESSING

QUICK SPARK OVERVIEW

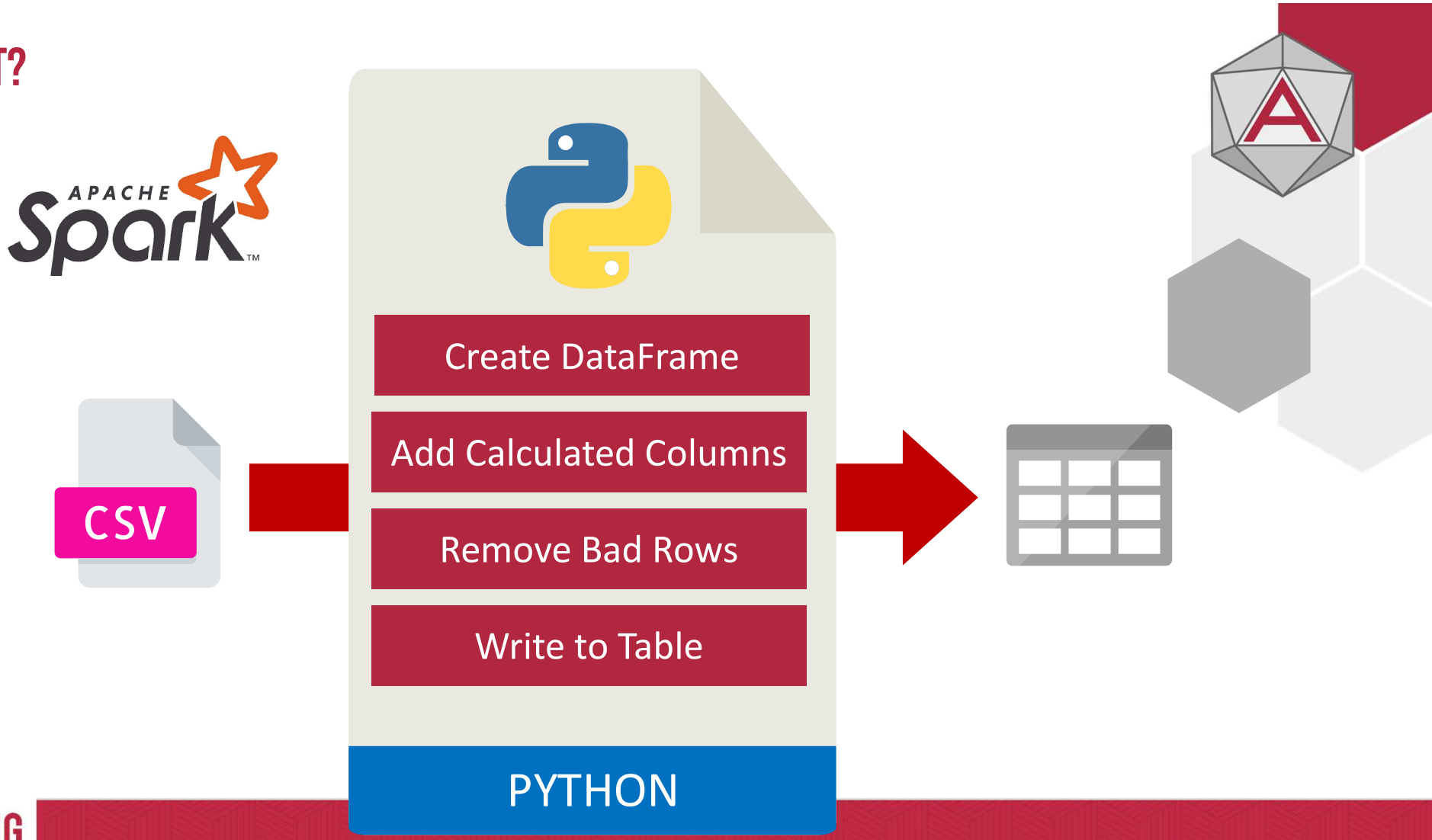
Spark is a distributed, scalable data processing engine.

It can query **structured** and **non-structured** data

You can use **Python**, **Scala**, **R**, **C#** or **SQL** to interact with it



SO WHAT?


APACHE
Spark
Create DataFrame
Add Calculated Columns
Remove Bad Rows
Write to Table
CSV
PYTHON

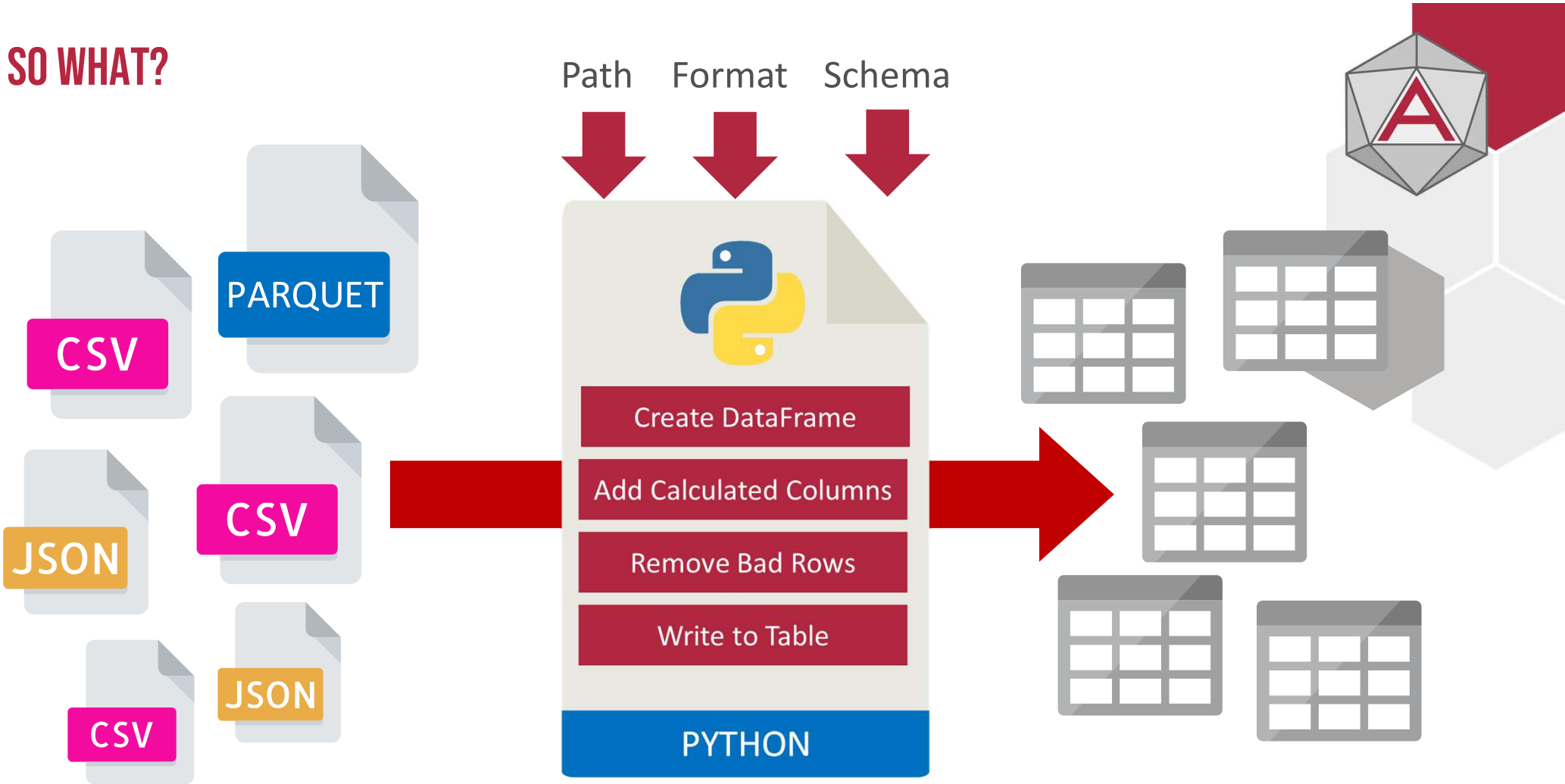
INSIDE A DATAFRAME

DataFrame

- Schema
- Format
- Location

```
df = (spark
        .read
        .schema(newSchema)
        .format(fileFormat)
        .load(dataLocation)
      )
```

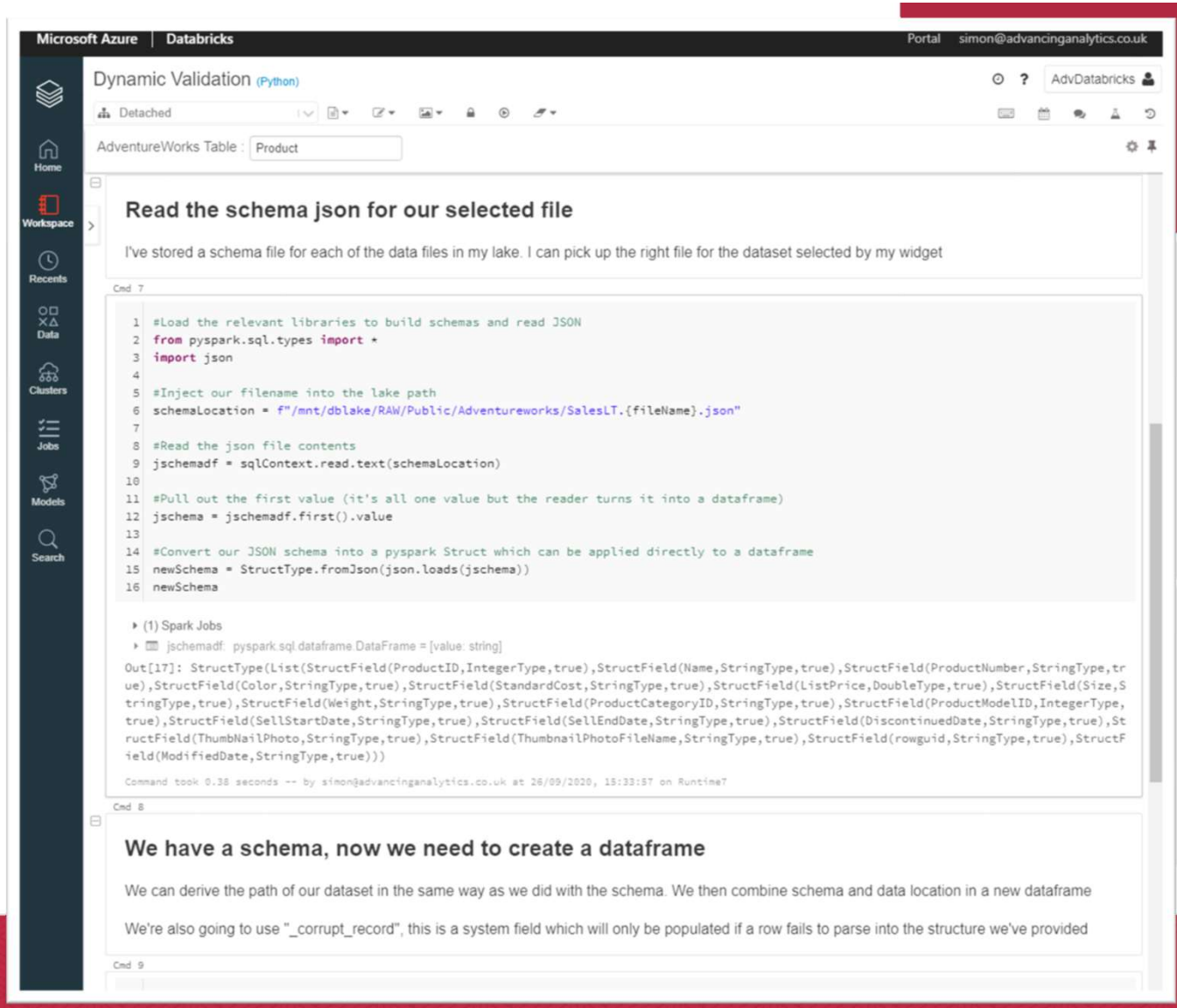
SO WHAT?


Path
Format
Schema
PARQUET
CSV
CSV
JSON
CSV
JSON
Create DataFrame
Add Calculated Columns
Remove Bad Rows
Write to Table
PYTHON

www.advancinganalytics.co.uk

SPARK OPTIONS IN AZURE

AZURE DATABRICKS



ADVANCING ANALYTICS

AZURE DATABRICKS

Databricks is a third-party company, founded by the **team who invented spark**

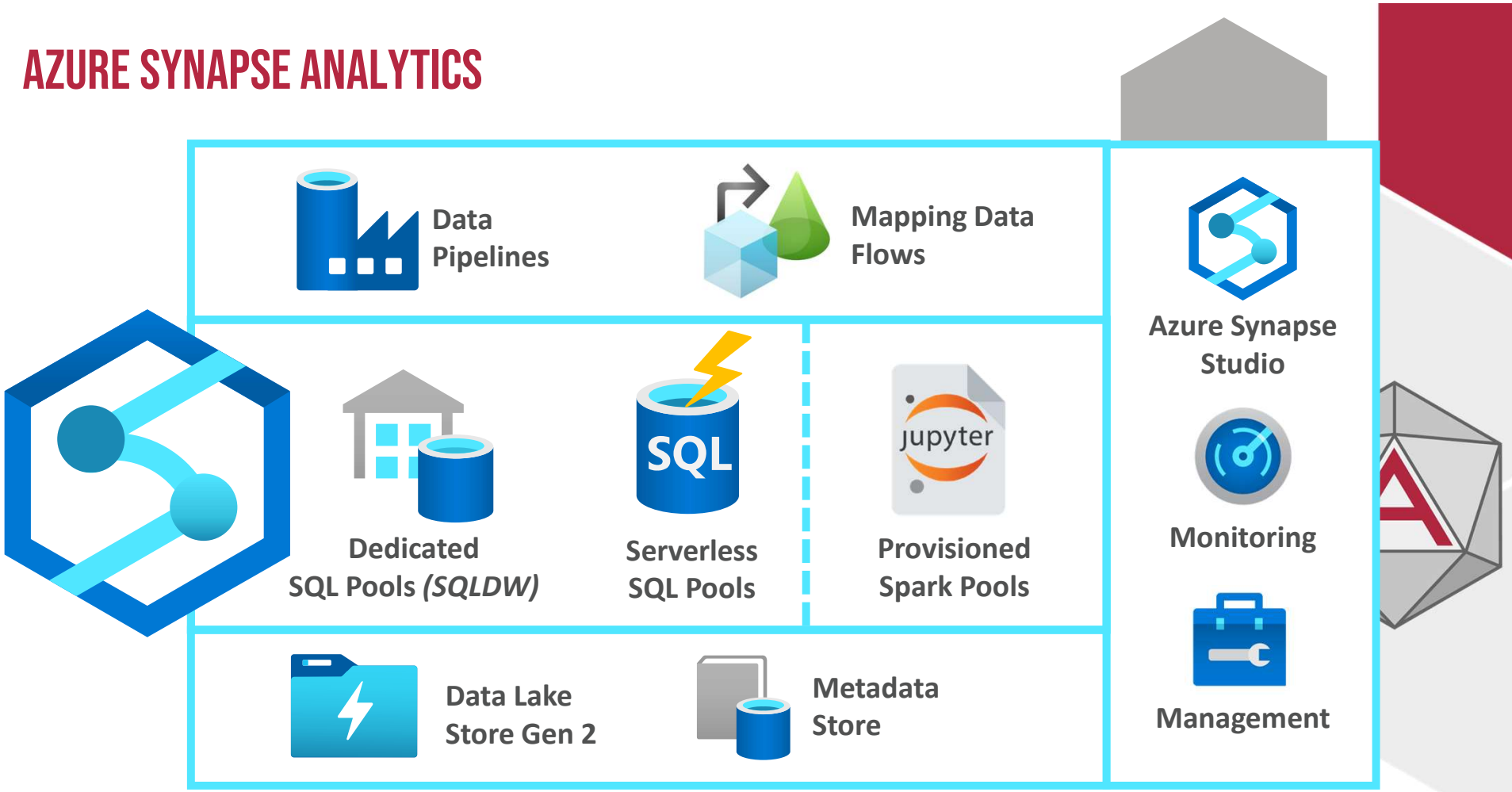
They provide an **Azure-native, managed Spark platform**

Databricks will generally have the **most advanced spark engine** and maintain a fast release cadence

AZURE DATABRICKS

- The Databricks Workspace
- Clusters
- Notebooks
- Libraries

AZURE SYNAPSE ANALYTICS



AZURE SYNAPSE ANALYTICS - SPARK POOLS

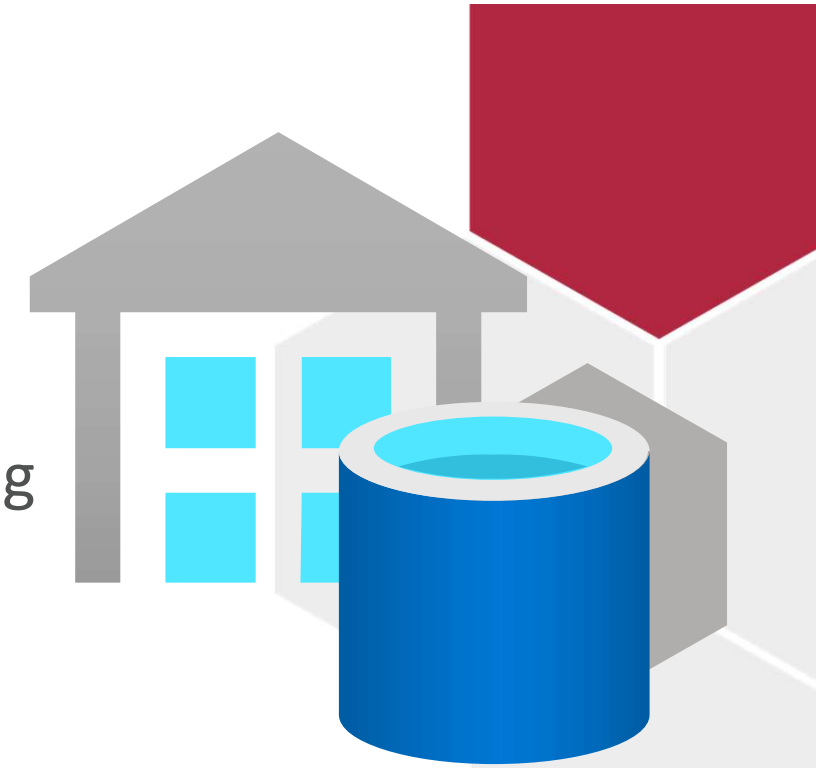
- Billed Per Session Uptime
- Scala, Python, C#, SQL
- Dynamic Workflows, Machine Learning & Unusual Data Types

SERVERLESS SQL POOLS

- Billed Per TB Read
- T-SQL
- Ad-hoc/Occasional access

DEDICATED SQL POOLS

- Billed Per Hour
- T-SQL
- Huge Datasets & Formal Modelling



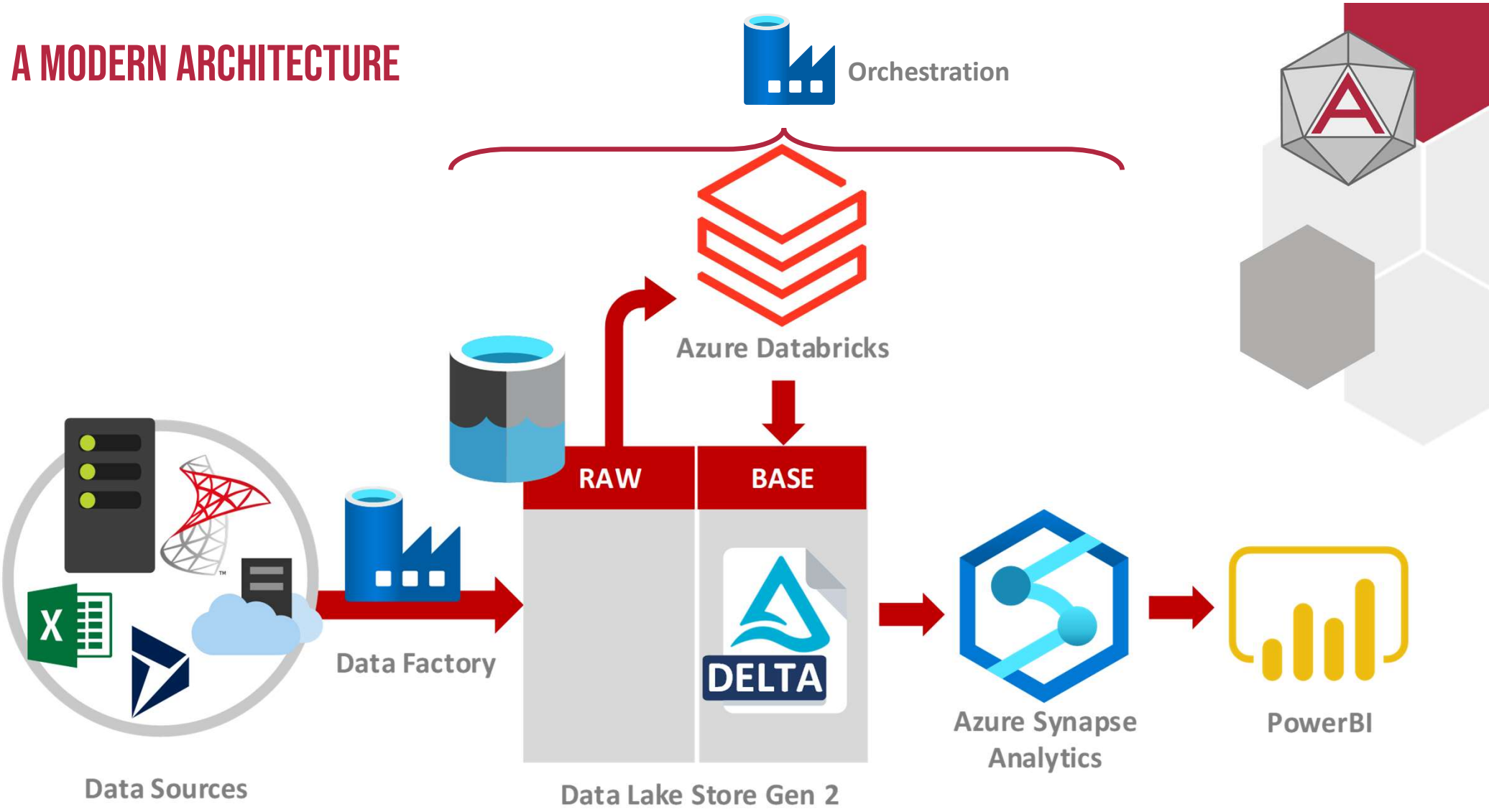
ADVANCING ANALYTICS

AZURE SYNAPSE ANALYTICS

- Synapse Workspace Overview
- Serverless SQL
- Spark Pools

SO WHERE DOES THAT LEAVE US?

A MODERN ARCHITECTURE



WHAT IS DATA ENGINEERING

How we Got Here

THE TOOLS AND TECHNOLOGY

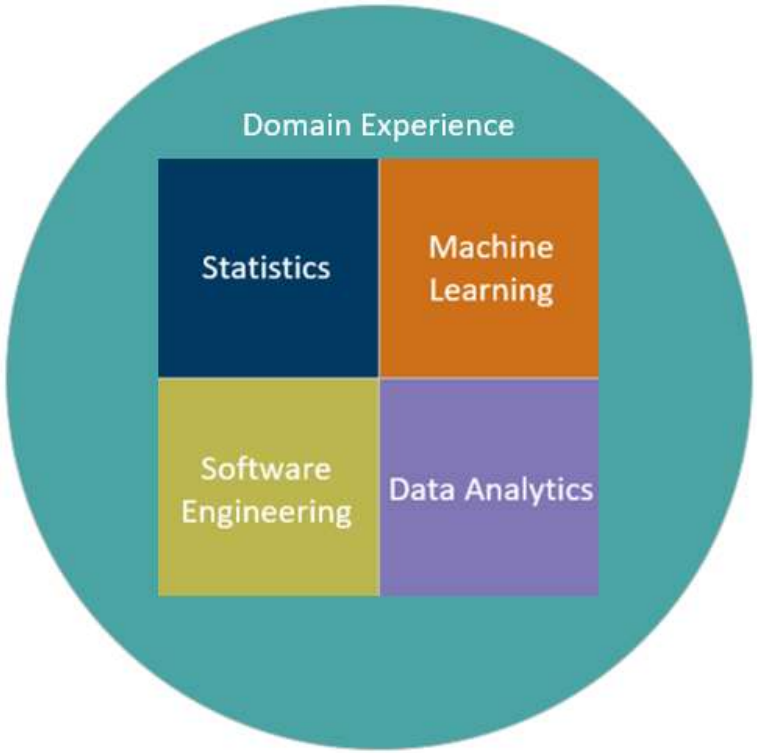
www.advancinganalytics.co.uk

DATA ENGINEERING VS DATA SCIENCE

The two work hand in hand

A **good Data Engineering Strategy** means less data wrangling for **Data Scientists**

DATA SCIENCE TEAMS

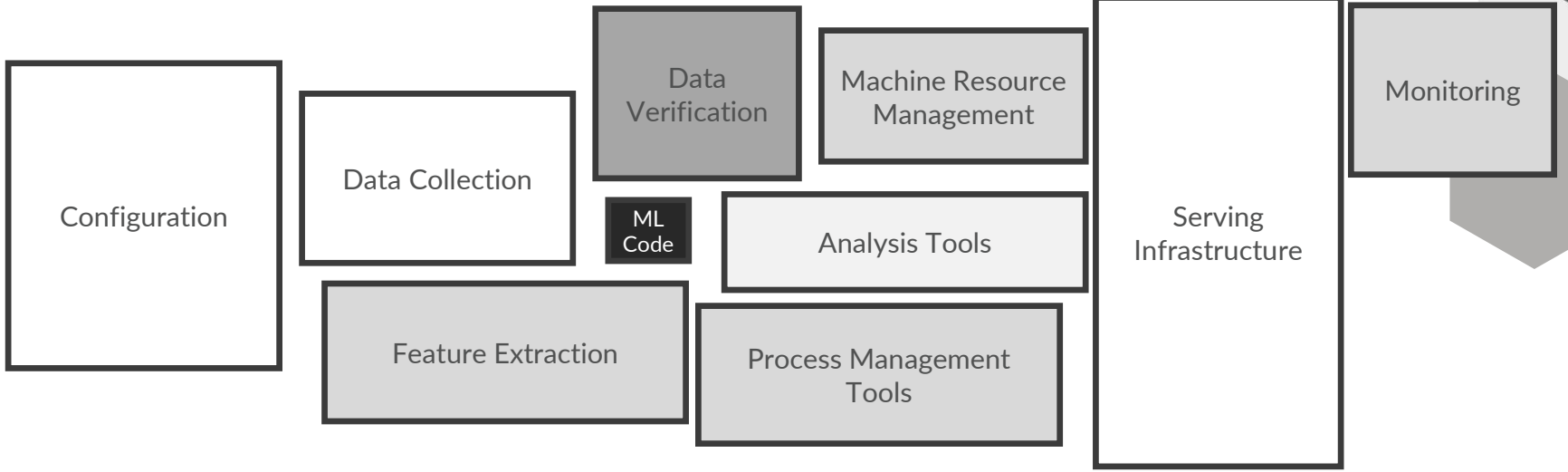


- STATISTICS
- MACHINE LEARNING
- SOFTWARE ENGINEERING
- DATA ANALYTICS

CRITICAL TO THE SUCCESS OF ANY MACHINE LEARNING

IS THAT THE BUSINESS DOMAIN IS WELL DEFINED AND UNDERSTOOD BY THE DEVELOPMENT TEAM.

HIDDEN DEBT IN ML SYSTEMS



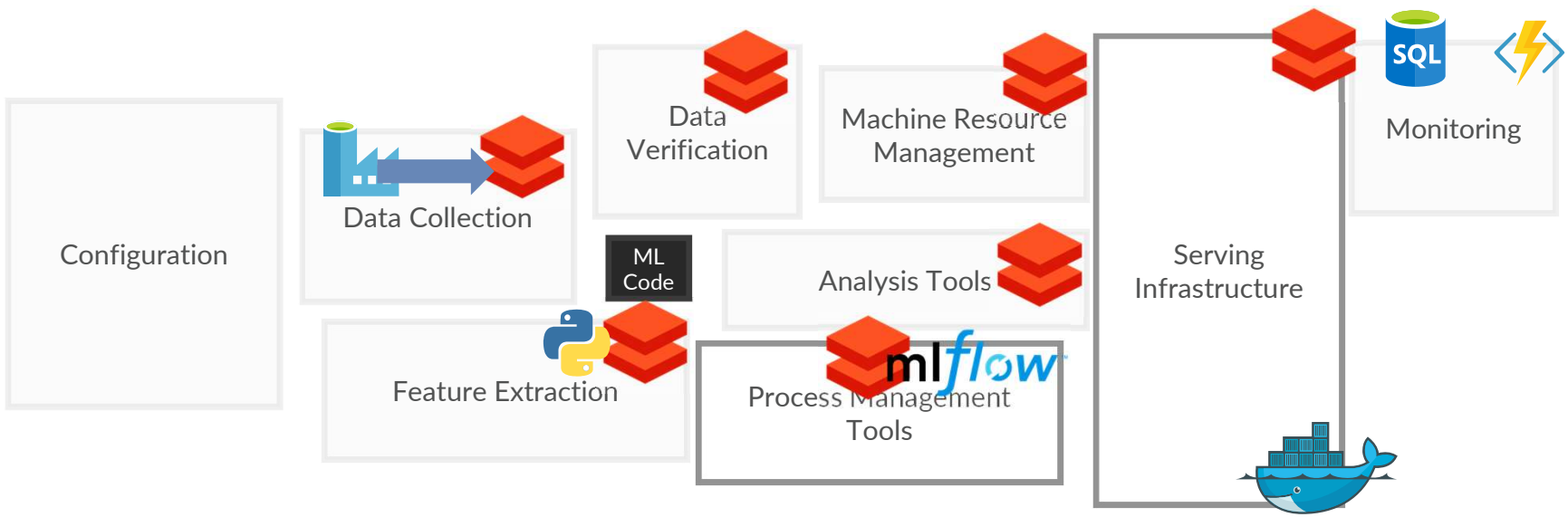
ML Code

Only a small amount of the time spent on a Machine Learning project is spent writing Machine Learning code. The rest of the time is spent on supplementary tasks.

80% of Data Science is spent on the rest of this diagram. After applying the common Data Engineering approaches we can reduce this

Scully et al, 2018

HIDDEN DEBT IN ML SYSTEMS



After implementing common engineering patterns the complexity in most areas are reduced. This can be reduced further with a standard approach to Machine learning deployments

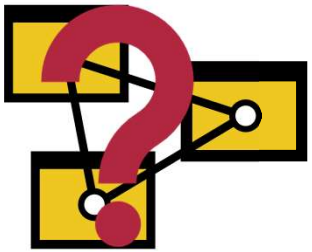
Scully et al, 2018

www.advancinganalytics.co.uk

DATA ENGINEERING VS BI

THE SKILL SETS CAN BE VERY SIMILAR, SO WHAT'S THE DIFFERENCE?

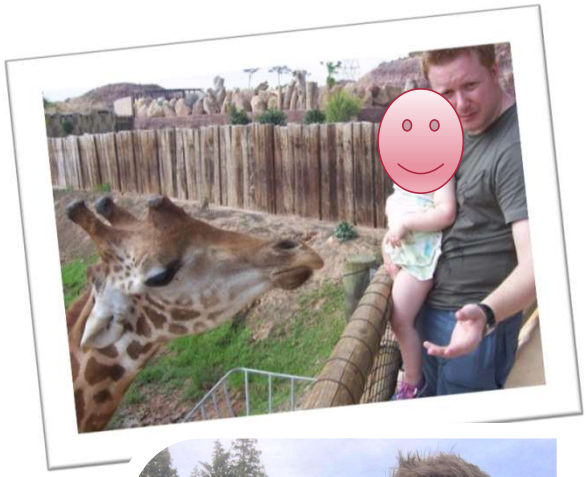
Data Engineers need to understand data modelling but don't necessarily have to implement data models:

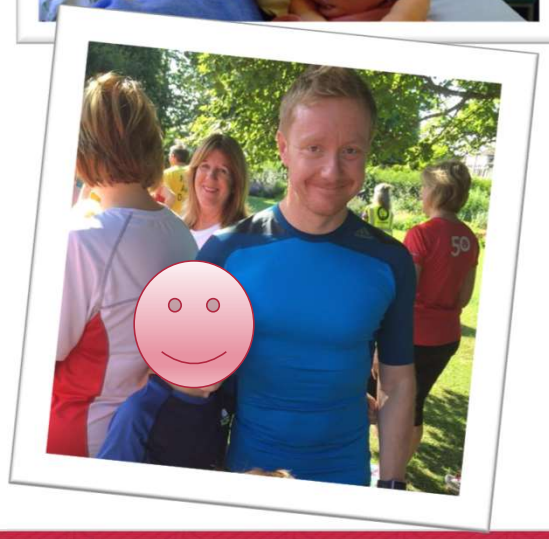


www.advancinganalytics.co.uk

DBA'S JOURNEY

MEET MIKEY... AS A DBA





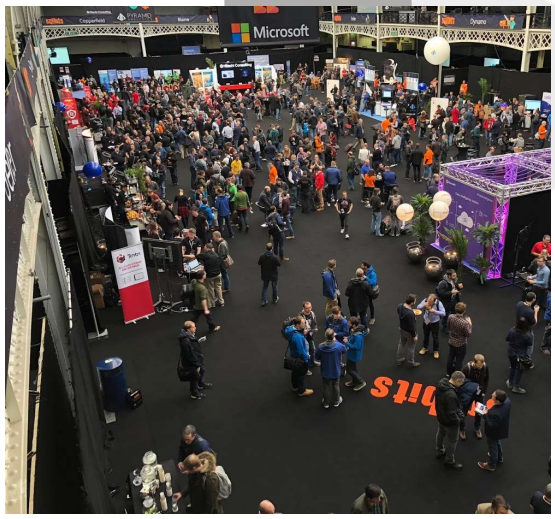
ADVANCING ANALYTICS

MY DBA JOURNEY AS EXPLAINED BY A DATA ENGINEER...

RAW

BASE

ENRICHED



CURATED

THEN...

THE JOURNEY

Server Admin

- Windows NT
- Netware
- Unix
- Security
- System maintenance
- Networks
- Storage

Accidental DBA

- PL SQL
- Shell Scripting

SQL DBA

- SQL
- Backup and Recovery
- Security
- Data Integrity
- SSIS
- SSAS
- SSRS

Senior DBA

- SAAS, PAAS, IAAS
- Performance tuning
- Process Definition & Ownership
- Architecture

SAP BI

- Warehousing
- Data Integration
- Data Modelling
- Reporting: Power Bi Business Objects
- Automation
- ETL
- Stakeholder Engagement
- SAP HANA

Data Engineering Consultant

- Python
- SQL
- Spark
- Azure
- Streaming
- Orchestration
- Agile & Scrum
- Strategy
- Architecture

ADVANCING ANALYTICS

www.advancinganalytics.co.uk

DATA ENGINEER VS DBA

WHAT YOU MAY BE USING TODAY

DBA's are heavily invested in Data and your skills reflect this

Relational

Microsoft SQL Server

ORACLE

Cloud

aws

Azure

Services

Azure Data Factory

Azure Data Lake Storage Gen2

SQL

Microsoft SQL Azure

Programming

SQL

python

WHAT CHANGES

DBA

- Most DBA's will work with relational platforms
- High priority on the incoming data being correct shape and size
- Thinking operationally how you manage the systems

DATA ENGINEER

- Data from and stored in structured and no-structured formats
- How we make data useable
- Think **parallel**, will it **scale**

SOFT SKILLS YOU MIGHT NOT KNOW YOU HAVE

- You can talk to tech's
- Engage with business
- Explain why

ADVANCING ANALYTICS

www.advancinganalytics.co.uk

A SOFTWARE ENGINEERS JOURNEY

WHERE IS ALL BEGAN

THE JOURNEY

Lab Technician (photography)	Graphic Artist	Application Developer	Software Developer/ Engineer	Data Engineer	Data Engineering Consultant
Hanging around in dark places	Adobe Software	XML JavaScript HTML C# SQL (Relational Databases) *Warehousing JSON Scala YML NoSQL Databases Kafka	XML JavaScript HTML C# SQL (Relational Databases) *Warehousing JSON Scala YML NoSQL Databases Kafka	XML JavaScript HTML C# SQL (Rela Database *Wareh JSON Scala YML NoSQL Da Kafka	Spark BigQuery Synapse Kimball Kappa Delta Lambda Architectures Agile & Scrum Python DataOps and MLOps

ADVANCING ANALYTICS

SOFTWARE ENGINEER TO DATA ENGINEER

- Already had a lot of the core skillsets required
- Used to working with CI (Continuous Integration)/CD (Continuous Deployment), IaC (Infrastructure as Code)
- Brought Software standards and processes into Data Engineering

www.advancinganalytics.co.uk

DATA ENGINEERING VS SOFTWARE ENGINEERING

THE SKILL SETS CAN BE VERY SIMILAR, SO WHAT'S THE DIFFERENCE?

OO Programmers are expected to know about Design Patterns, Functional Programmers are expected to know about Category Theory.

So on the same note a Data Engineer should know about the different Data processing and storage architectures, including:

- Kappa
- Delta
- Lambda architectures
- Star Schemas

ADVANCING ANALYTICS

THE SKILL SETS CAN BE VERY SIMILAR, SO WHAT'S THE DIFFERENCE?

Data Engineers need to understand different Data store offerings, and when to use them, including:

Relational

Microsoft SQL Server

ORACLE

Key Value

Amazon DynamoDB

redis

Graph

neo4j

Azure Cosmos DB

Document

elastic

mongoDB.

THE SKILL SETS CAN BE VERY SIMILAR, SO WHAT'S THE DIFFERENCE?

Data Engineers need to know which compute offering to use and when, and where to host them including:

Kubernetes

kubernetes

docker

Custom Application

Azure web app

node

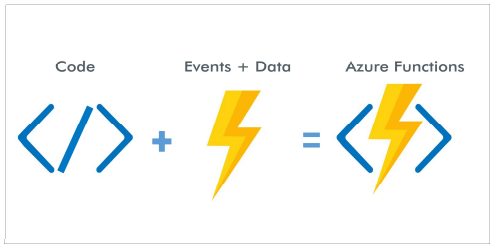
Spark/Databricks

databricks

APACHE Spark

Event Driven Serverless Functions

Amazon Lambda



www.advancinganalytics.co.uk

A DATA ENTHUSIAST'S JOURNEY

A LONG TIME AGO...

THE JOURNEY

Army

Getting cold, wet (often miserable) in lovely places…

Data Analyst

SPSS
Python
Excel (lots of Excel)

Insight Manager

Excel (even more Excel)
SPSS
SQL
R

BI Consultant

SQL
SSIS
SSAS
SSRS
Power BI
Warehousing
C#
Agile & Scrum
Architecture

Head of BI

Management
Strategy
Warehousing
Excel (SO MUCH!!)
PowerPoint
Teams
SQL
SSAS
Power BI
Datazen??

Data Engineering Consultant

Python
SQL
Spark
Excel
PowerPoint
Power BI
Agile & Scrum
Strategy
Architecture

DATA GENERALIST TO DATA ENGINEER

- Already had a lot of the core skillsets required
- Used to working with ETL and BI Tools
- Brought end-user perspective to Engineering

HOW ARE WE GOING TO HELP

- WHAT IS DATA ENGINEERING 
- THE TOOLS AND TECHNOLOGY 
- How we Got Here 

CALL TO ACTION

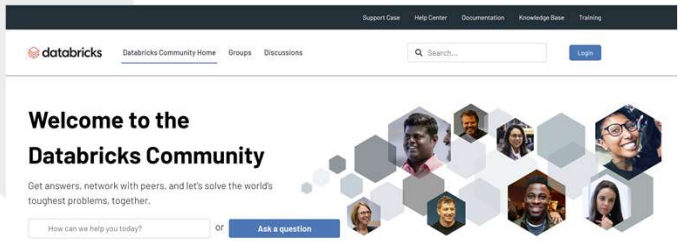
BOOKS YOU SHOULD DEFINITELY READ

ADVANCING ANALYTICS

WEBSITES FOR HELP AND LEARN

stackoverflow

cathrine WILHELMSEN



Spark by {Examples}

- www.cathrinewilhelmsen.net
- community.databricks.com
- sparkbyexamples.com
- stackoverflow.com

And of course

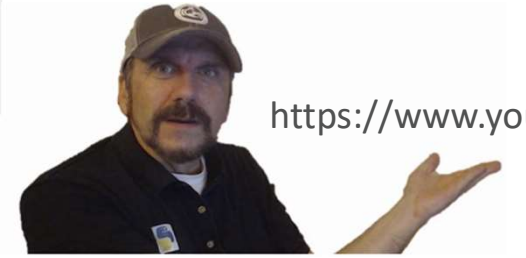
- www.advancinganalytics.co.uk

VIDEO LINKS

https://www.youtube.com/c/AdvancingAnalytics

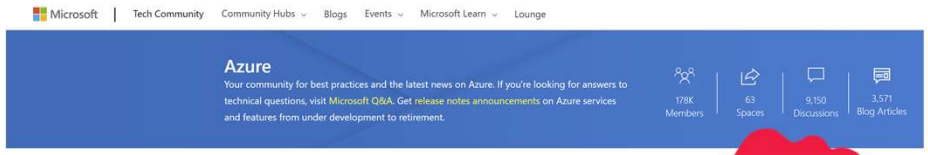
https://www.youtube.com/c/SeattleDataGuy

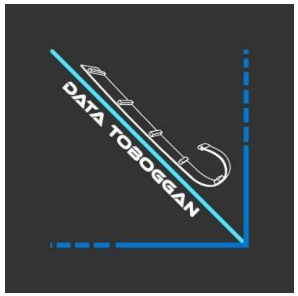
https://www.youtube.com/c/Databricks



https://www.youtube.com/c/BryanCafferky

USE THE COMMUNITY





meetup

DATA RELAY

sqlbits

SQLBITS SESSIONS COMING NEXT

Straight After this….

- Kevin Chant, Kay Cordewener - Azure Data Engineering services used to analyze data (Room 9)

- Just Blindbæk - Architectural blueprints for the Modern Data Warehouse (Room 12)

Then @ 17:10

- Kamil Nowinski - Azure Data Factory - A deployment challenges (Room 12)

- Craig Porteous, Chris Williams - "Cultivating the Catalogue" - Growing Data Governance with Azure Purview (Room 3)

- Cathrine Wilhelmsen, Helle Normann - Lessons Learned: Implementing Azure Synapse Analytics in a Rapidly-Changing Startup (Room 7)

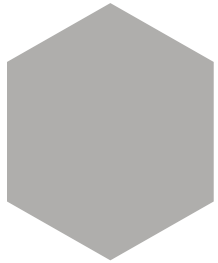
SQLBITS SESSIONS TOMORROW

- Ivan Donev - Practical Lakehouse implementation in the enterprise
- Heini Ilmarinen - Less Clicking, More Coding! Data Platform Development Using Infrastructure as Code
- Mike Dobing - Building the Lakehouse Architecture with Azure Synapse Analytics
- Angela Henry - What is Azure Purview and Why Do I Need It?
- Zach Stagers - Synapse Data Flows - Will Citizen ETL Replace the Data Engineer?
- Falek Miah, Anna Wykes - Automate the deployment of Databricks components using Terraform
- Bernhard Engleder, Wolfgang Strasser - Data Governance with Azure Purview - Theory, Customer Insights and Demos
- Dustin Vannoy - Azure Data Engineer Skills for Success
- Oliver Engels, Tillmann Eitelberg - Enterprise Scale Analytics - What's in it for me
- Koen Verbeeck - The modern Cloud Data Warehouse - Snowflake on Azure
- Ben Jarvis - Azure Container Apps for Data Engineers
- Ajith Ramanath - Hitting Top Speeds with Spark SQL in Azure Synapse
- Piotr Mucha - Run your SSIS packages in Azure Data Factory

SQLBits Sessions
Grab the Recordings

- Luke Moloney - The fundamentals of building a lakehouse with Synapse
- André Kamman - Building your first Metadata Driven Azure Data Factory
- Nick Baladi, Dave Draffin, Martyn Bullerwell,Andrew Isenman - Heathrow's Data Transformation, How to rapidly build out a data engineering function
- Simon Whiteley - Lessons in Lakehouse Automation
- Andy Cutler - The Dream Team: Synapse Analytics Serverless SQL Pools and Pipelines
- Mike Diehl - Azure DevOps Release Pipelines for SQL Databases, Azure Data Factories, and Analysis Models
- Jean Joseph - From Housekeeping to Data Engineer - My journey to find my passion
- Niels Berglund - ksqlDB - The Real-Time Streaming Database
- Simon Whiteley - Bringing Data Lakes to your Purview
- Gavin Campbell, Edward Allen - Data Engineering the Hard Way: Developing Spark Applications on Kubernetes
- James Serra - Data Lakehouse, Data Mesh, and Data Fabric (the alphabet soup of data architectures)
- Paul Andrew - Creating a Metadata Driven Orchestration Framework Using Azure Data Integration Pipelines

SQLBITS SESSIONS
GRAB THE RECORDINGS



- Kevin Chant, Sander Stad - Github and Azure DevOps Duet - SQLBits edition
- Ust Oldfield - Implementing a Data Quality Framework in Purview
- Daniel Harrington - Leveraging Apache Spark for Efficient Data Encryption at Scale
- Henk van der Valk - Lessons learned from managing Azure Data platforms! Put your data hat on!
- Paul Andrew - The Evolution of Data Platform Architectures in Azure - Lambda, Kappa, Delta, Data Mesh
- Evangeline White, Gaurav Malhotra - Unified Data Governance with Azure Purview
- Erwin de Kreuk - Lake Database with Database Template and Mapping Data with Azure Synapse Analytics
- Gaurav Malhotra - Self Service Data Access Provisioning With Azure Purview
- Matthew Roche - Roche's Maxim of Data Transformation
- Vanessa Araújo - Spring clean your data lake with Database Templates for Synapse Analytics
- Dr. Victoria Holt, Erwin de Kreuk, Wolfgang Strasser - Data Governance with Azure Purview - Ask the Experts

UST OLDFIELD

PRINCIPAL ADVANCING ANALYTICS CONSULTANT SPECIALIZING IN **DATA ENGINEERING, ANALYTICS** AND **STRATEGY.**

OVER 10 YEARS' EXPERIENCE, MOST RECENTLY WORKING WITHIN THE **MICROSOFT DATA PLATFORM.**

DATA AND TECH ENTHUSIAST – LOOKING TO MAKE DATA ACCESSIBLE TO ALL

UST-OLDFIELD

@USTDOESTECH

ANNA WYKES

SENIOR ADVANCING ANALYTICS CONSULTANT SPECIALIZING IN **DATA ENGINEERING, DEVOPS** & **CLOUD**.

OVER 16 YEARS' EXPERIENCE, MOST RECENTLY WORKING WITHIN THE **MICROSOFT DATA PLATFORM**, **SCALA**, **KAFKA** AND VARIOUS CLOUD TECH

HELP TO RUN LOCAL DATA MEETUP "**DATA BRISTOL**", AND ALSO HELP OUT AT LOCAL CODE CLUBS

BSC IN MULTIMEDIA COMPUTING & BUSINESS, AND A HND IN VISUAL COMMUNICATION

ANNA-MARIA-WYKES

@ANNAWYKES

Microsoft® MVP Most Valuable Professional

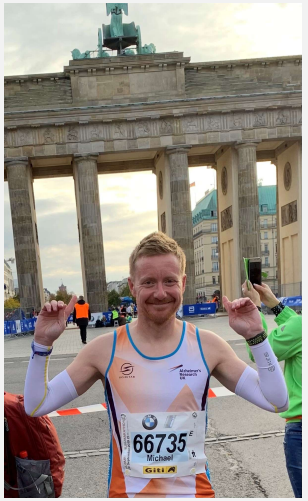
MICHAEL ROBSON

SENIOR ADVANCING ANALYTICS CONSULTANT SPECIALIZING IN **DATA ENGINEERING, ANALYTICS** AND **STRATEGY**.

THIS ~~IS~~ WAS MY FIRST PRESENTING SESSION, ANYWHERE!

WAS A DBA FOR 20 YEARS

BEEN INVOLVED IN THE DATA COMMUNITY FOR OVER 10 YEARS, SETUP DATA PLATFORM MEETUP "**NEWCASTLE DPAC**", HELP ORGANISE DATA RELAY AND HAVE BEEN HELPER AT SQL BITS SINCE 2015



MICHAELROBSONUK

@HEYMIKY

PLEASE LEAVE FEEDBACK

https://sqlb.it/?7129