Инжиниринг Данных в Microsoft

основано на реальных событиях

SmartData 2022

Disclaimer

All thoughts are mine and don’t represent Microsoft or any other company.

Содержание

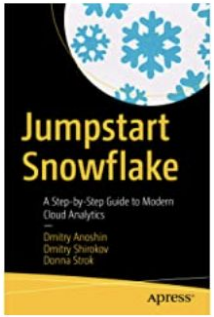
- Про меня
- Собеседование и начало работы;
- Создание Delta Lake с нуля и модернизация аналитического решения;
- специфика Microsoft;
- плюсы и минусы работы в Microsoft;
- отличие культур Amazon и Microsoft;
- типовые аналитические решения Microsoft.

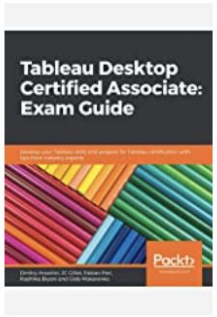
- 12 + лет в аналитики, до этого был инженером конструктором на ГКНПЦ им Хруничева
- Москва -> Черногория ->Канада
- 5 лет @Amazon, 2 года @Microsoft Gaming
- Tableau, Snowflake, Microsoft, AWS пользовательские группы и митапы, конференции

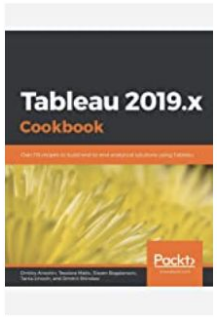
DataLearn.ru
7000 Students
- DE 101
- DS&ML 101
- SQL 101

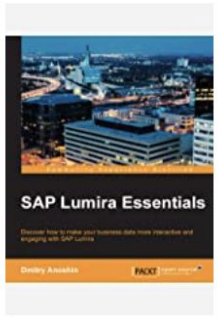
University of Victoria

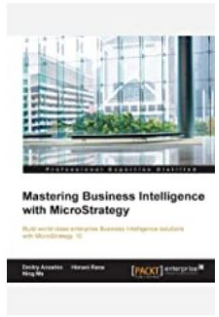
Инжиниринг данных
https://t.me/rockyourdata

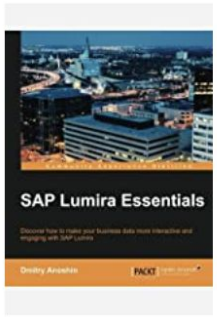








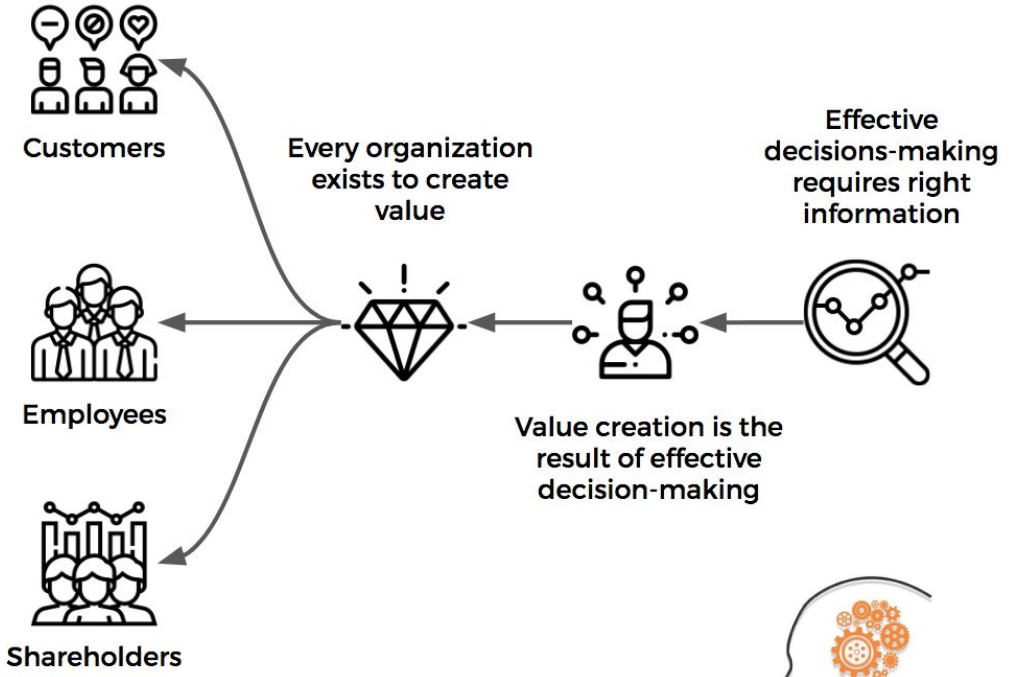


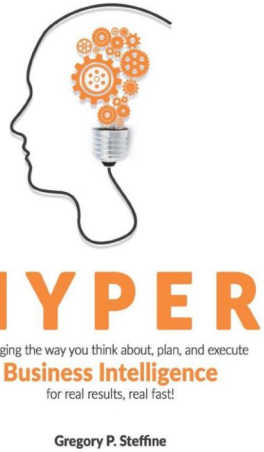


Часть 1: Немножко теории

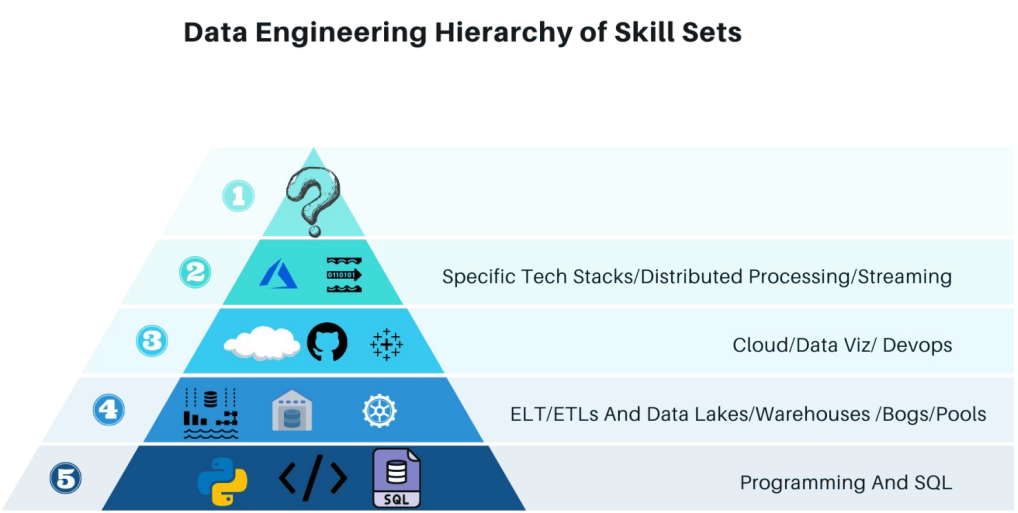
Зачем нужна аналитика?

- Повысить прибыль
- Сократить расходы
- Избежать риски
- Исследование новых рынков и продуктов
- Проверить гипотезы



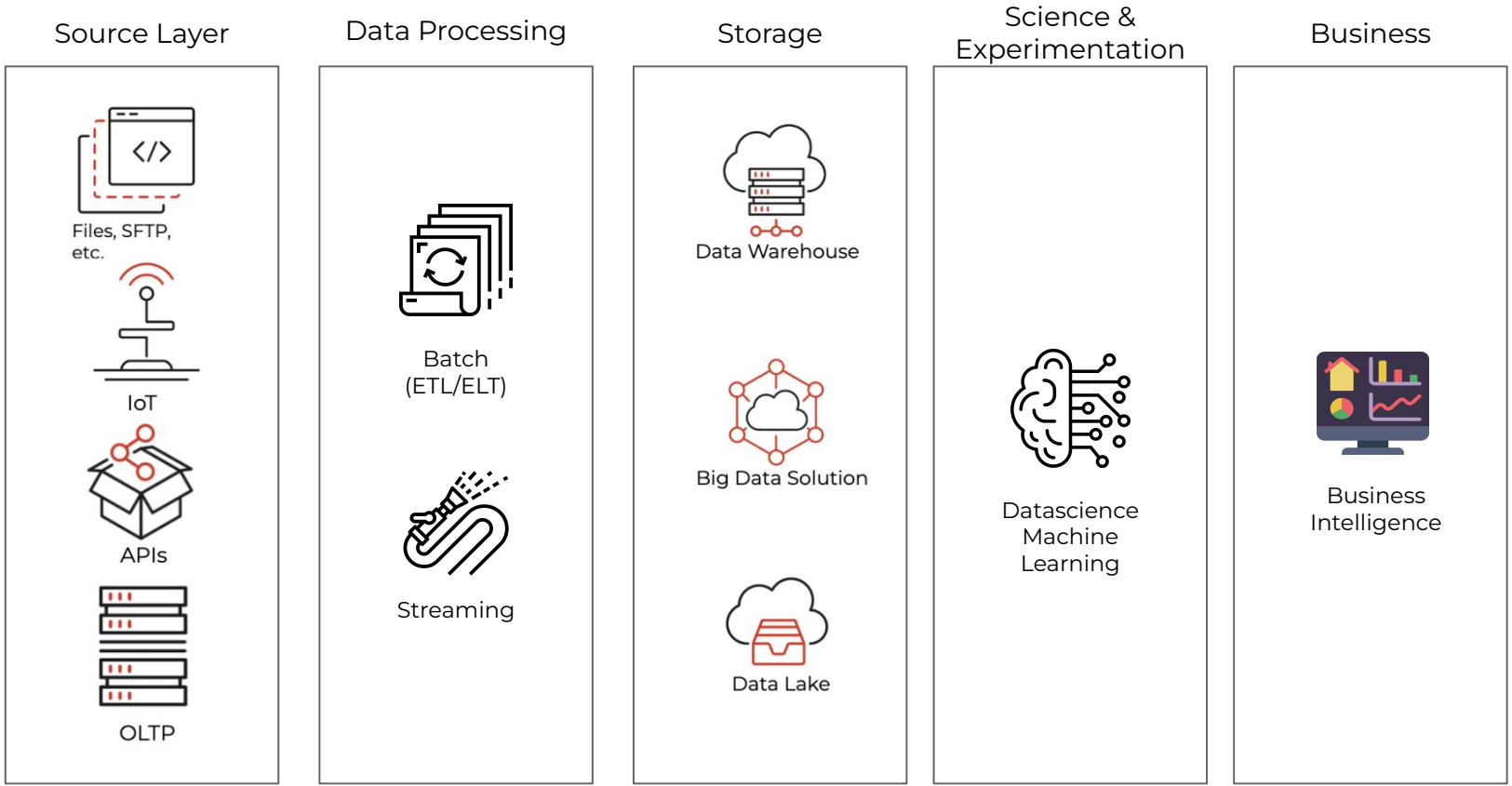


Что такое Инжиниринг данных?



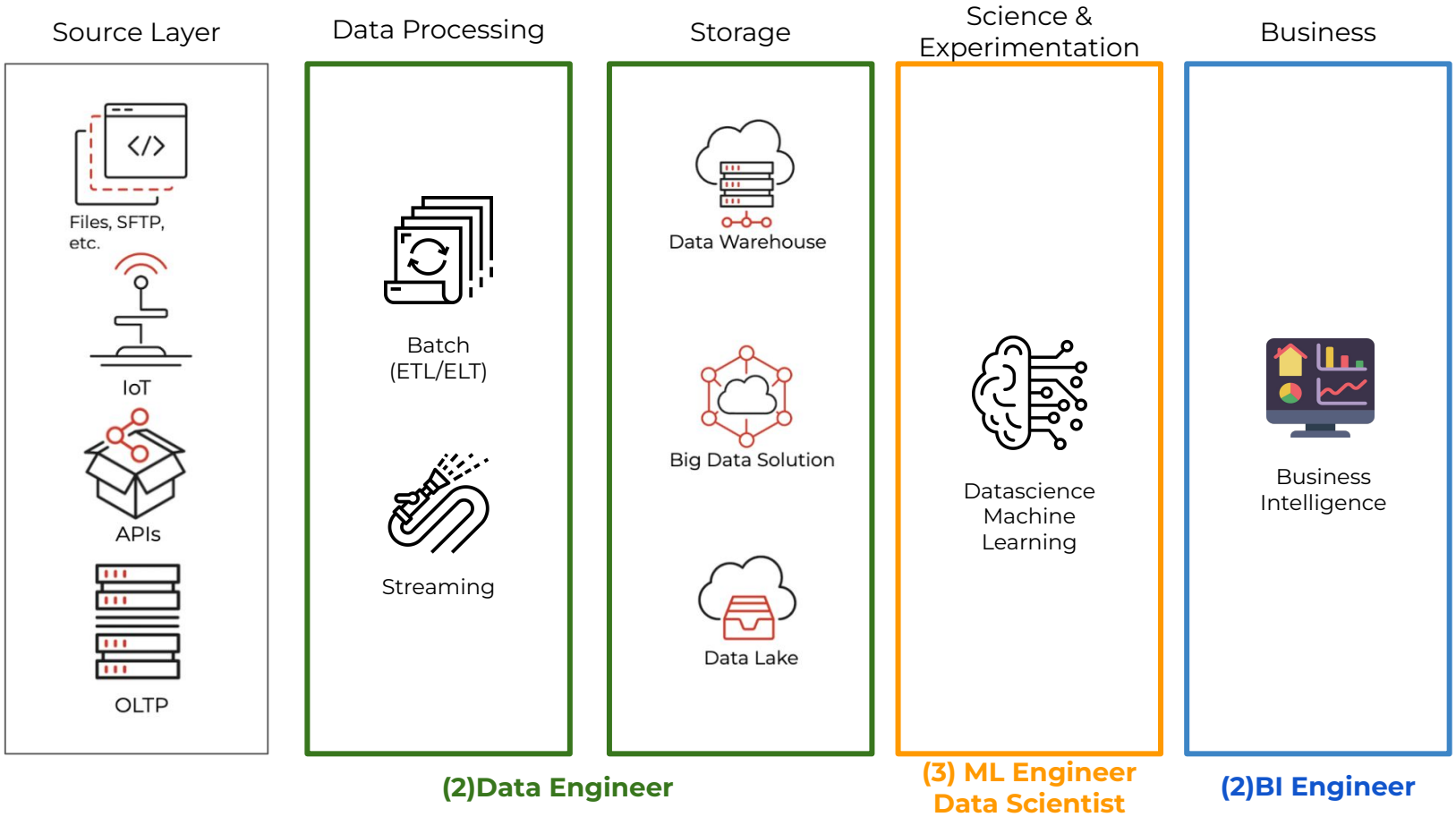
Инжиниринг данных делает данные **полезными** и **доступными** для потребителей, создавая **безопасную** и **масштабируемую** инфраструктуру данных.

~~Modern~~ Data Stack



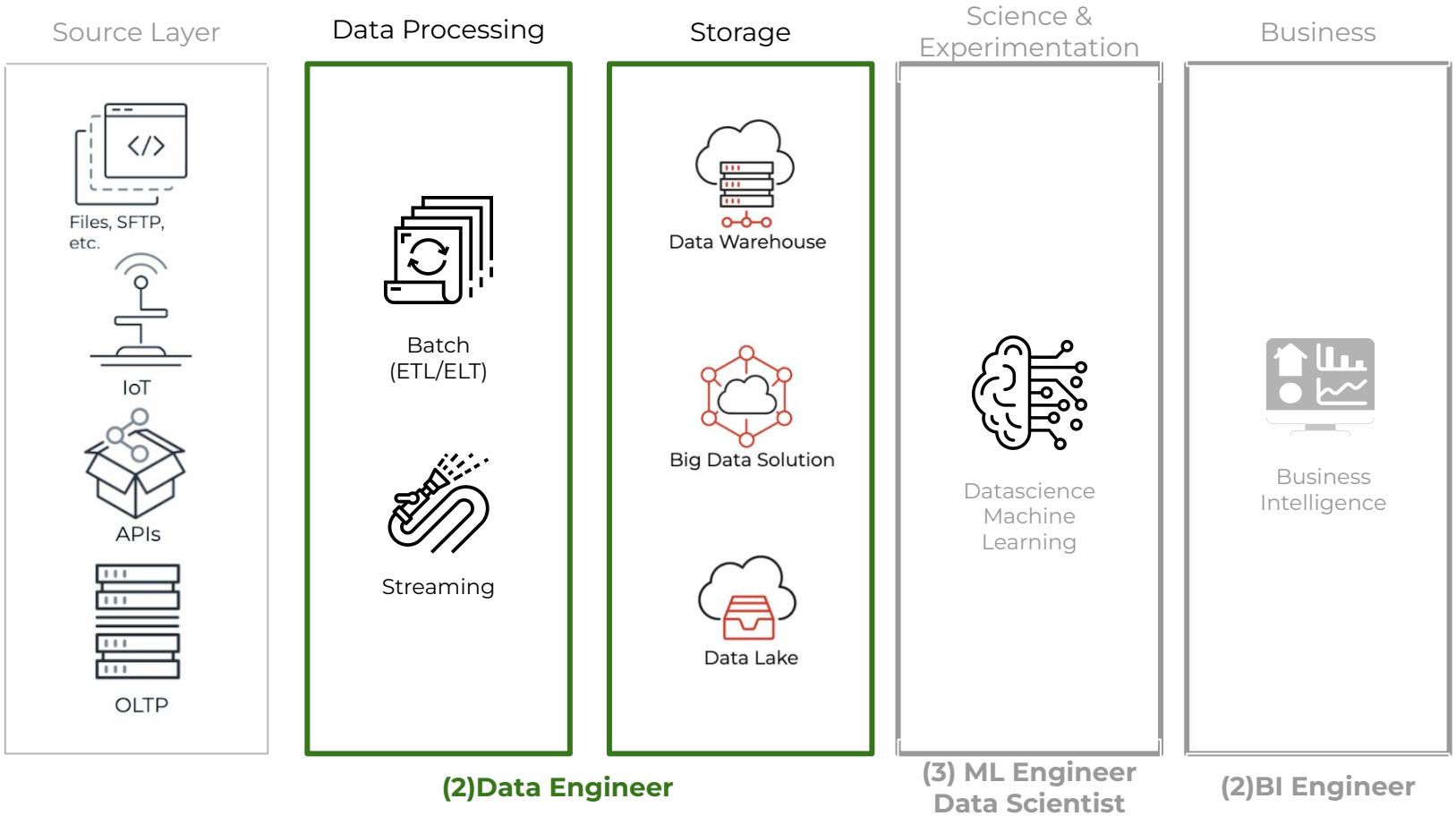
Key Layers and roles

(1)Product Manager - manage data product.



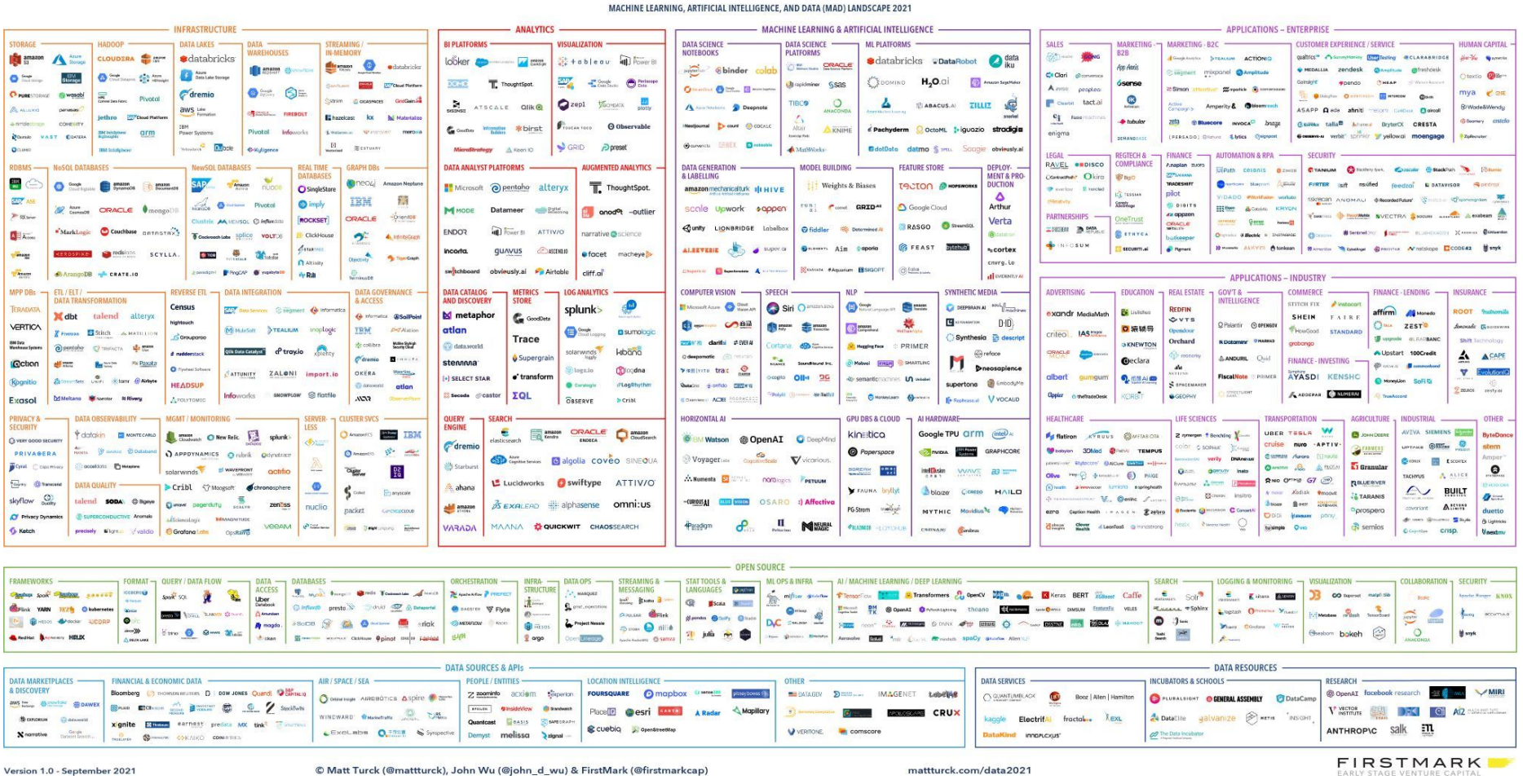
DE Key Layers

(1)Product Manager - manage data product.



With DE you can - “move fast, break things”(c)...

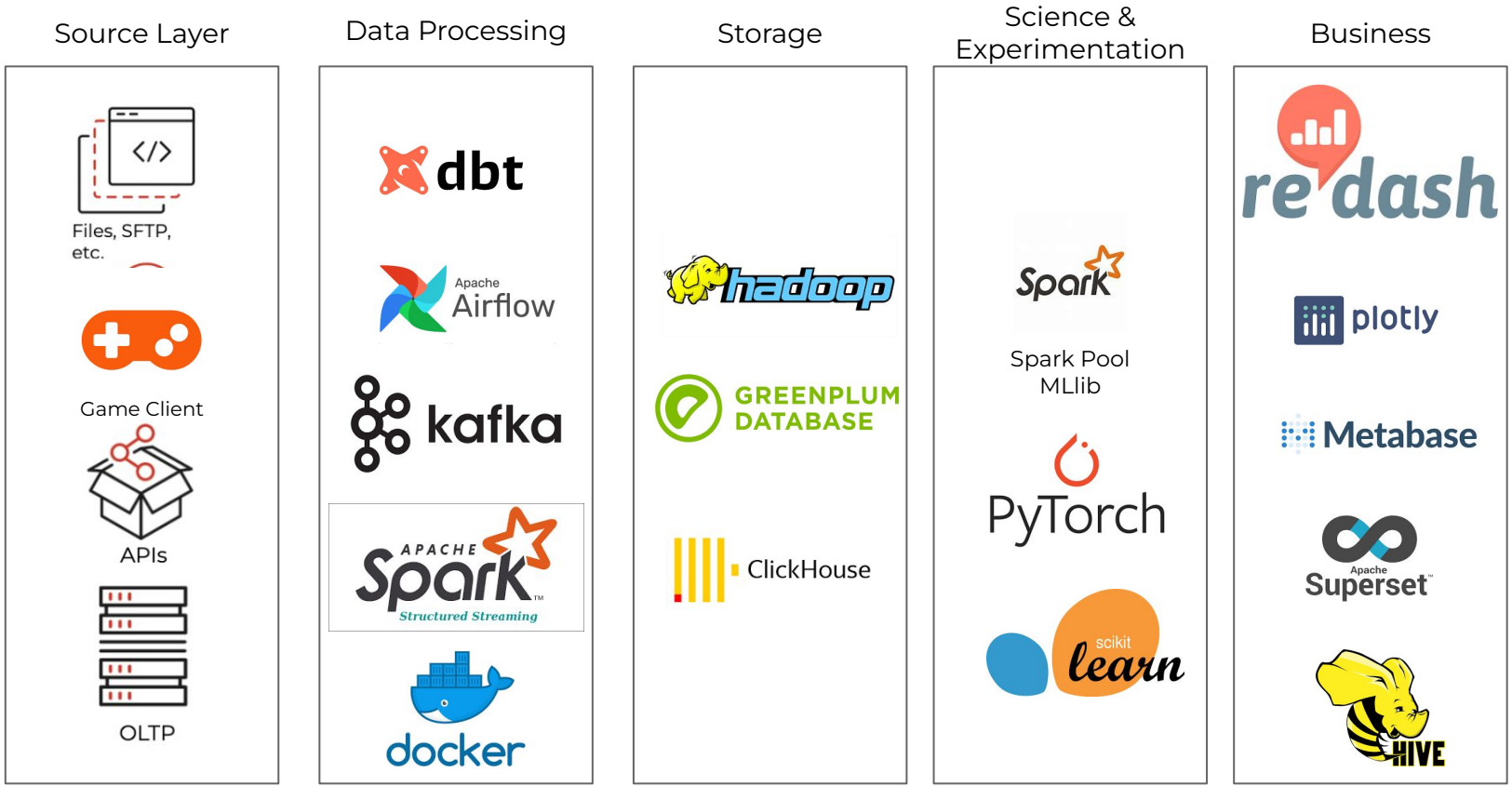
Data and AI Landscape 2021



© Matt Turck (@mattturck), John Wu (@john_d_wu) & FirstMark (@firstmarkcap)

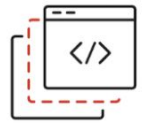
FIRSTMARK
EARLY STAGE VENTURE CAPITAL

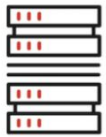
Data Stack with Open Source



Data Stack for $$$

Source Layer





Data Processing

Census

dbt_

Fivetran

MATILLION

hightouch

Storage

Google Big Query

snowflake

amazon REDSHIFT

databricks

Science & Experimentation

alteryx

DataRobot

Datascience
Machine
Learning

Business

looker

Business
Intelligence

Часть 2: Начало работы в Microsoft

От отклика до начала работы

- Отклик Май
- 1 й звонок конец июля
- собеседования Август-Сентября
- Начало работы Ноябрь

Итоге: 5 месяцев

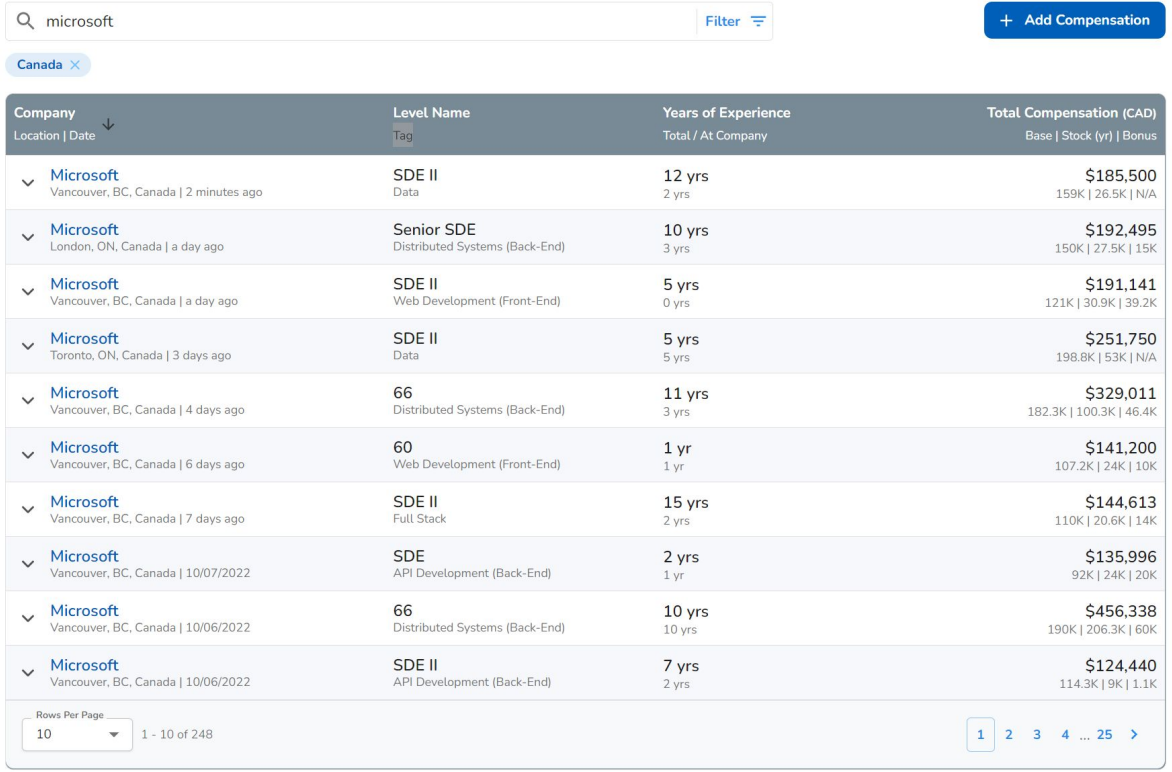
Уровни Amazon vs Microsoft

https://www.levels.fyi/?compare=Amazon,Microsoft&track=Software%20Engineer

Что входит в оффер

- Sign Up bonus 1 и 2-й год
- RSU (акции на период 4 года)
- 10-20% годового бонуса
- базовая зарплата
- страховка
- отчисления на пенсию
- премиум переезд и 2 месяца в центре города для семьи

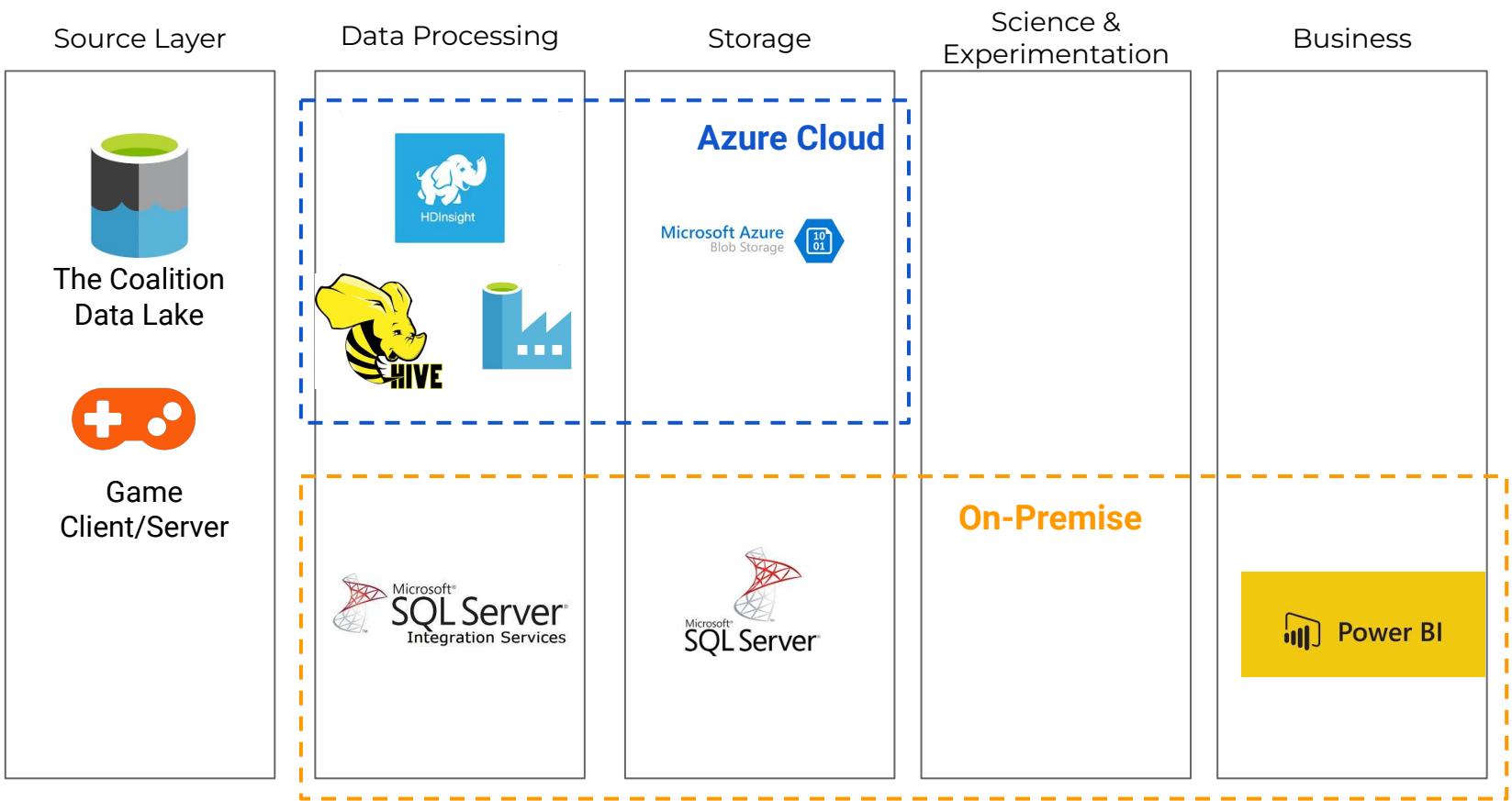
PS самый низкий total comp среди Faang, но хорошие бенефиты.



Company Location | Date	Level Name Tag	Years of Experience Total / At Company	Total Compensation (CAD) Base | Stock (yr) | Bonus
Microsoft Vancouver, BC, Canada | 2 minutes ago	SDE II Data	12 yrs 2 yrs	$185,500 159K | 26.5K | N/A
Microsoft London, ON, Canada | a day ago	Senior SDE Distributed Systems (Back-End)	10 yrs 3 yrs	$192,495 150K | 27.5K | 15K
Microsoft Vancouver, BC, Canada | a day ago	SDE II Web Development (Front-End)	5 yrs 0 yrs	$191,141 121K | 30.9K | 39.2K
Microsoft Toronto, ON, Canada | 3 days ago	SDE II Data	5 yrs 5 yrs	$251,750 198.8K | 53K | N/A
Microsoft Vancouver, BC, Canada | 4 days ago	66 Distributed Systems (Back-End)	11 yrs 3 yrs	$329,011 182.3K | 100.3K | 46.4K
Microsoft Vancouver, BC, Canada | 6 days ago	60 Web Development (Front-End)	1 yr 1 yr	$141,200 107.2K | 24K | 10K
Microsoft Vancouver, BC, Canada | 7 days ago	SDE II Full Stack	15 yrs 2 yrs	$144,613 110K | 20.6K | 14K
Microsoft Vancouver, BC, Canada | 10/07/2022	SDE API Development (Back-End)	2 yrs 1 yr	$135,996 92K | 24K | 20K
Microsoft Vancouver, BC, Canada | 10/06/2022	66 Distributed Systems (Back-End)	10 yrs 10 yrs	$456,338 190K | 206.3K | 60K
Microsoft Vancouver, BC, Canada | 10/06/2022	SDE II API Development (Back-End)	7 yrs 2 yrs	$124,440 114.3K | 9K | 1.1K

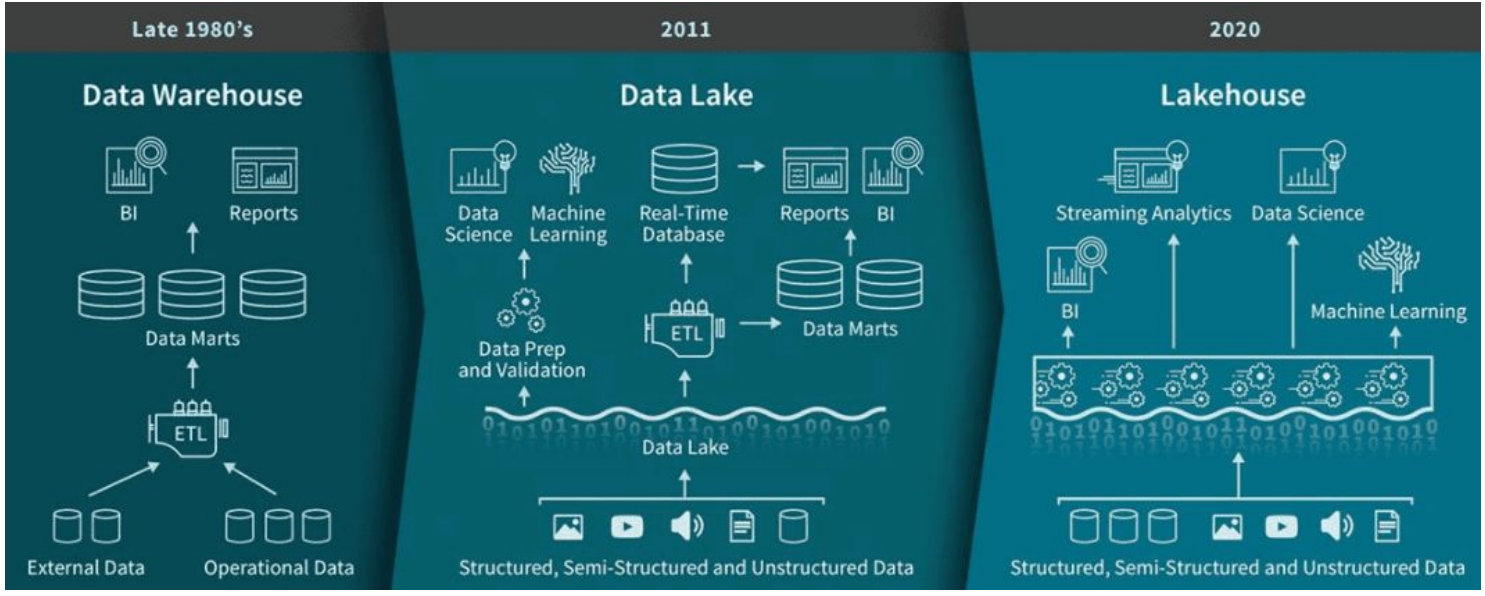
Часть 3: Модернизация и миграция

Было: HDInsights + Hive (staging), SQL Server SSIS (fact tables)



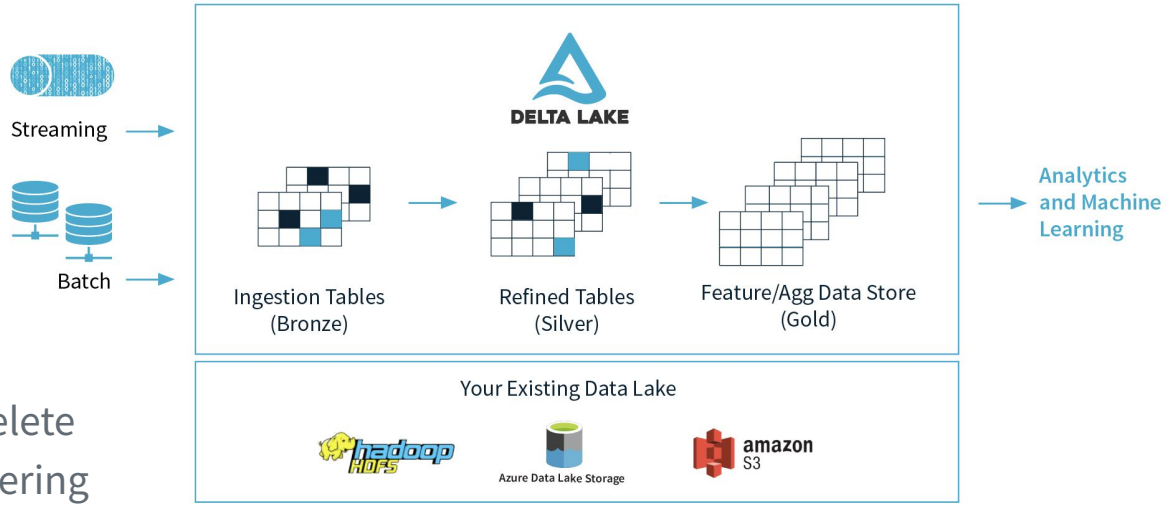
Analytics architecture evolution

- До 2010 использовали хранилища данных (SMP, MPP).
- После появления Hadoop, все бросились строить озера данных. Преимущества - раздельный слой хранения и обработки данных, но нет поддержки ACID.
- Lake house = условно хранилище + озеро данных (Snowflake, Databricks, Synapse)



Lake House = DW + Data Lake

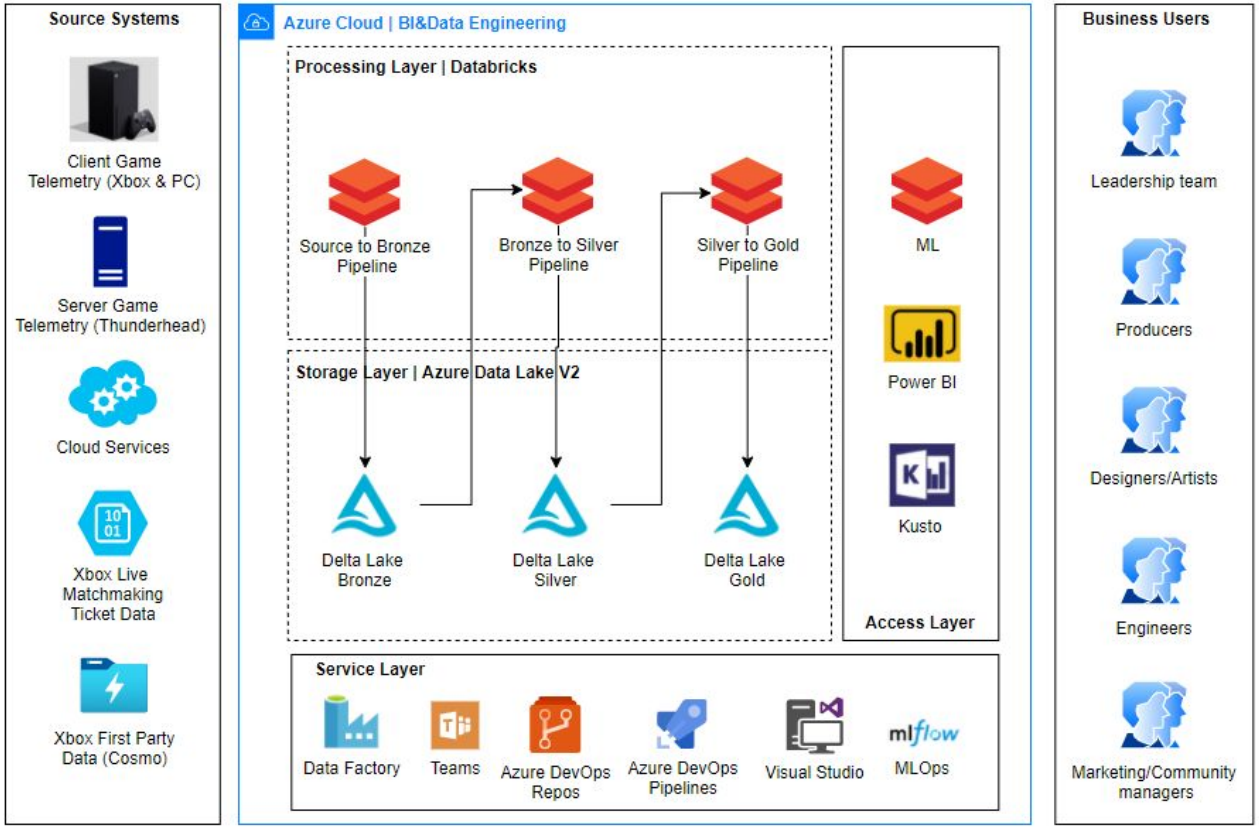
- Transaction Support (ACID)
- Schema Enforcement
- Upserts/Deletes



Key solutions for Lakehouse:

- **Apache Hudi** (Hadoop Update Delete and Incremental) by Uber Engineering
- **Apache Iceberg** by Netflix
- **Delta Lake** by Apache Spark

Стало: Databricks+ Delta Lake



Задача Инженера Данных в Gaming

- создание инфраструктуры данных для сбора телеметрии
- содание сред разработки, тестирования и продакшн
- трансформации бизнес логики в PySpark/SQL/C#
- автоматизация data pipelines
- качество данных, документация
- подключение к BI системам
- участие в разработки исходного формата и типа телеметрии
- тестирование игры

Сложности:
- во время разработки (2-3) года у вас не ПРОД объема данных
- кол-во уникальных событий сотни, практически каждое событие имеет свою уникальную метрику

Telemetry as a source of Player data

The word Telemetry is derived from the Greek roots tele, "remote", and metron, "measure".

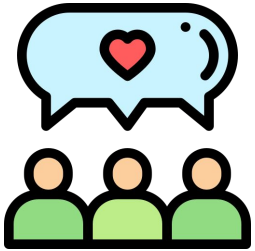
Games are **state machines** - a person creates a **continual loop** of actions and responses which keep the game state changing. Often loops keeping the user **engages over a period of time**.

Telemetry helps to discovering **who** is performing **what** action **when** and **where** in the game. It cannot provide **why**.

3 types of metrics

Gameplay metrics
user behavior in the game

Customer metrics
user as a customer, acquisition and retention

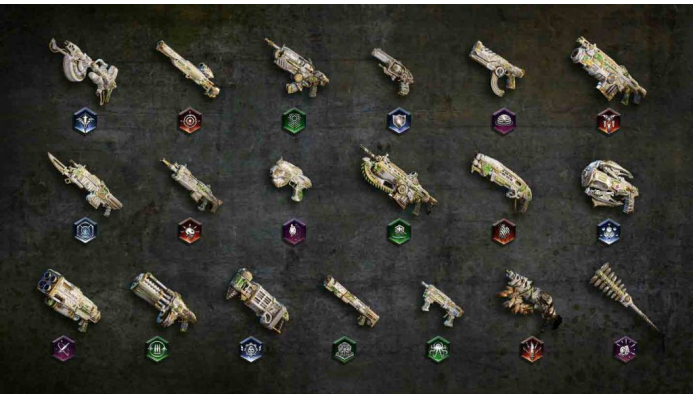


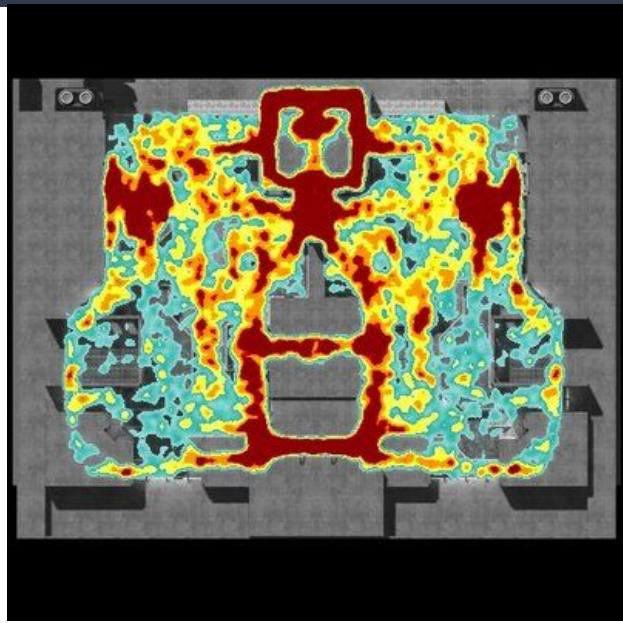
Community metrics
user engagement in communities and social media

Action Third-Person Shooters (TPS) Metrics

- Weapon use
- trajectory
- item/asset use
- character/kit choice
- level/map choice
- loss/win
- heatmaps

- team scores
- map lethality
- map balance
- vehicle use metrics
- special moves
- jumps and many more.





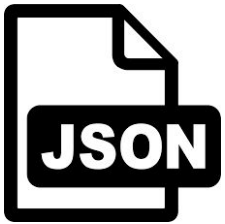
Death map, Halo3

https://coolinfographics.com/blog/2009/1/12/halo-3-heatmaps.html

Data Engineering Design Flow as a Funnel

Event Names:

- Weapon Use
- Damage
- Shooting
- Flock
- Map Name
- HeartBeat
- and so on



Raw Tables (Bronze)
Method: Append
Trans: Minimum

Staging Tables (Silver)
Method: Append
Trans: JSON Schema

Fact Tables (Gold)
Method: Merge
Trans: Heavy

Databricks custom logs

Monitoring Azure Databricks in an Azure Log Analytics Workspace

This repository extends the core monitoring functionality of Azure Databricks to send streaming query event information to Azure Monitor. For more information about using this library to monitor Azure Databricks, see Monitoring Azure Databricks

The project has the following directory structure:

```
/src
    /spark-listeners-loganalytics
    /spark-listeners
    /pom.xml
/sample
    /spark-sample-job
/perftools
     /spark-sample-job
```

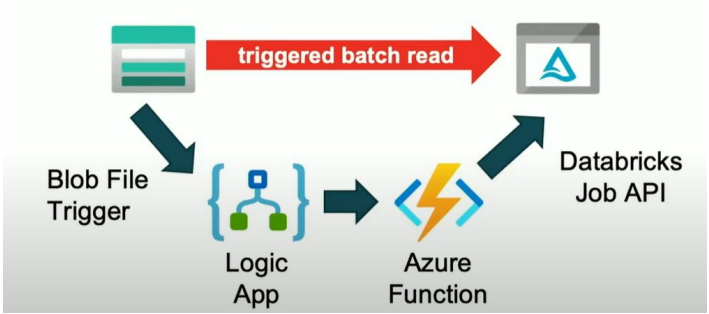
The **spark-listeners-loganalytics** and **spark-listeners** directories contain the code for building the two JAR files that are deployed to the Databricks cluster. The **spark-listeners** directory includes a **scripts** directory that contains a cluster node initialization script to copy the JAR files from a staging directory in the Azure Databricks file system to execution nodes. The **pom.xml** file is the main Maven project object model build file for the entire project.

The **spark-sample-job** directory is a sample Spark application demonstrating how to implement a Spark application metric counter.

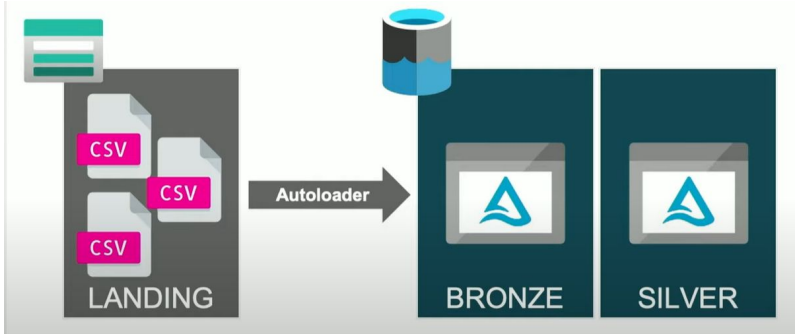
The **perftools** directory contains details on how to use Azure Monitor with Grafana to monitor Spark performance.

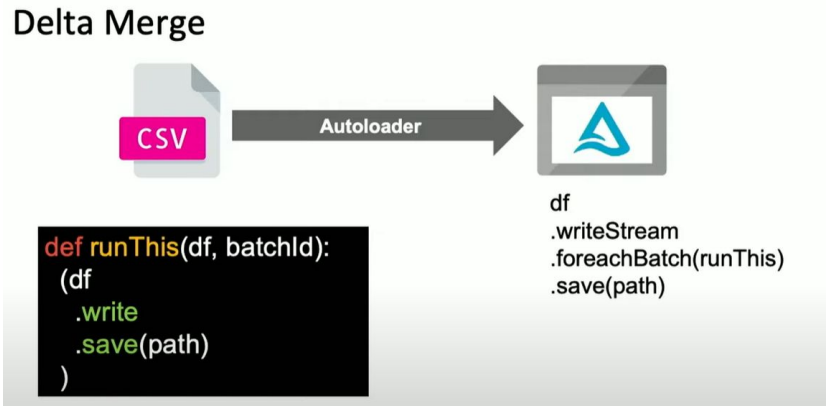
https://github.com/mspnp/spark-monitoring

Databricks Delta Streaming (Auto Loader)



As files are discovered, their metadata is persisted in a scalable key-value store (RocksDB) in the checkpoint location of your Auto Loader pipeline. This key-value store ensures that data is processed exactly once.





Из интересного

Пожаловался лично Филу, что не дают работать инженеру и дают использовать сырой продукт Synapse...

Проблемы была решена оперативно.

Phil Spencer named Microsoft Gaming CEO following Activision Blizzard deal

By Zachary Boddy published January 18, 2022

Microsoft's gaming leadership is uniting under Phil Spencer's guidance.

Comments (2)

Phil Spencer Xbox 20 Celebration (Image credit: Microsoft)

Часть 4: Специфика DE работы в Microsoft

Есть ли позиция Data Engineer в Microsoft?

А кто есть?

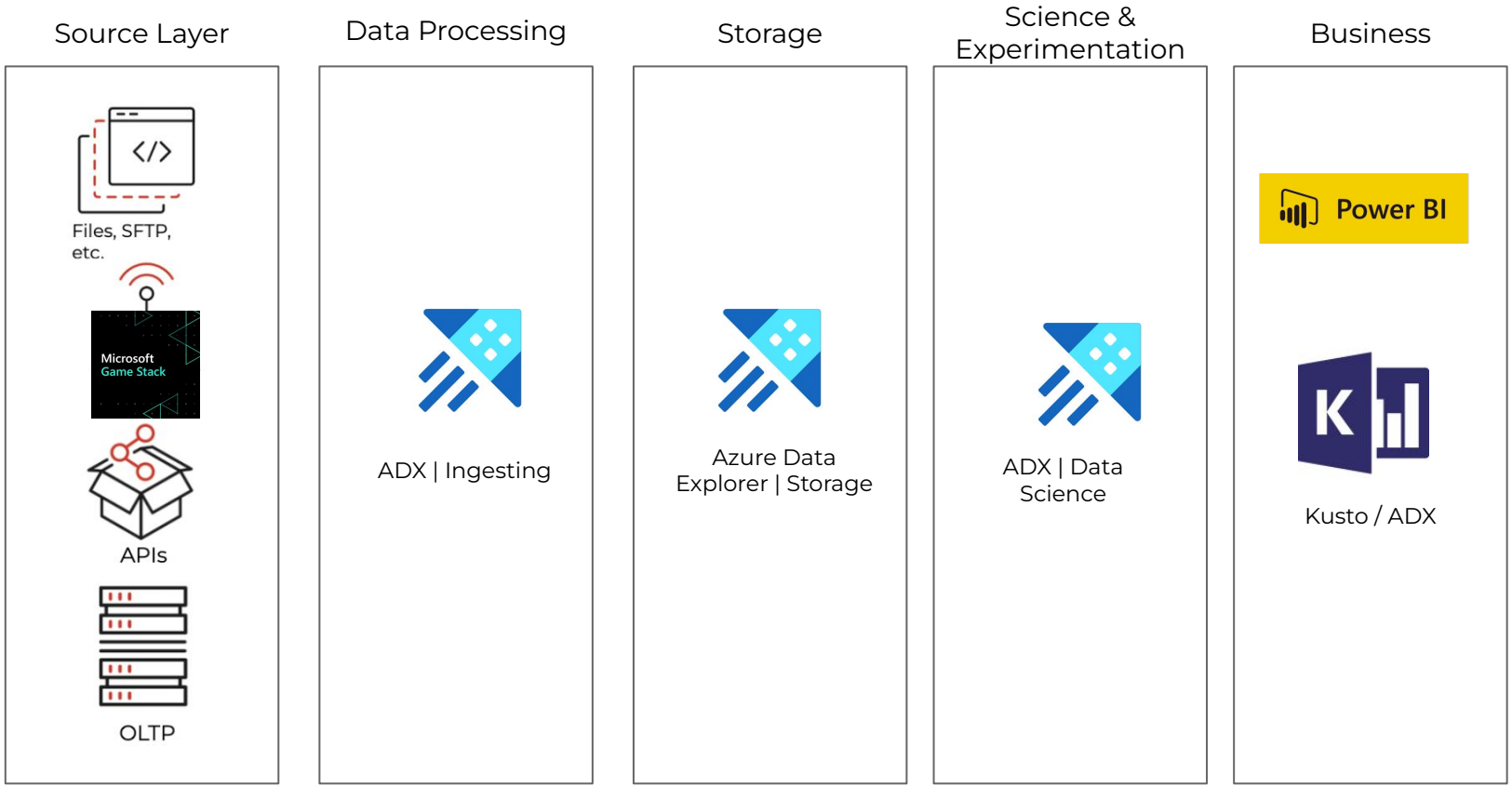
- Software Development Engineer (SDE и DE)
- Data & Applied Scientist (analyst, BI, data science)
- Business Analyst (может быть и data analyst)

В Амазоне для BI и DE (SDE все Оки)

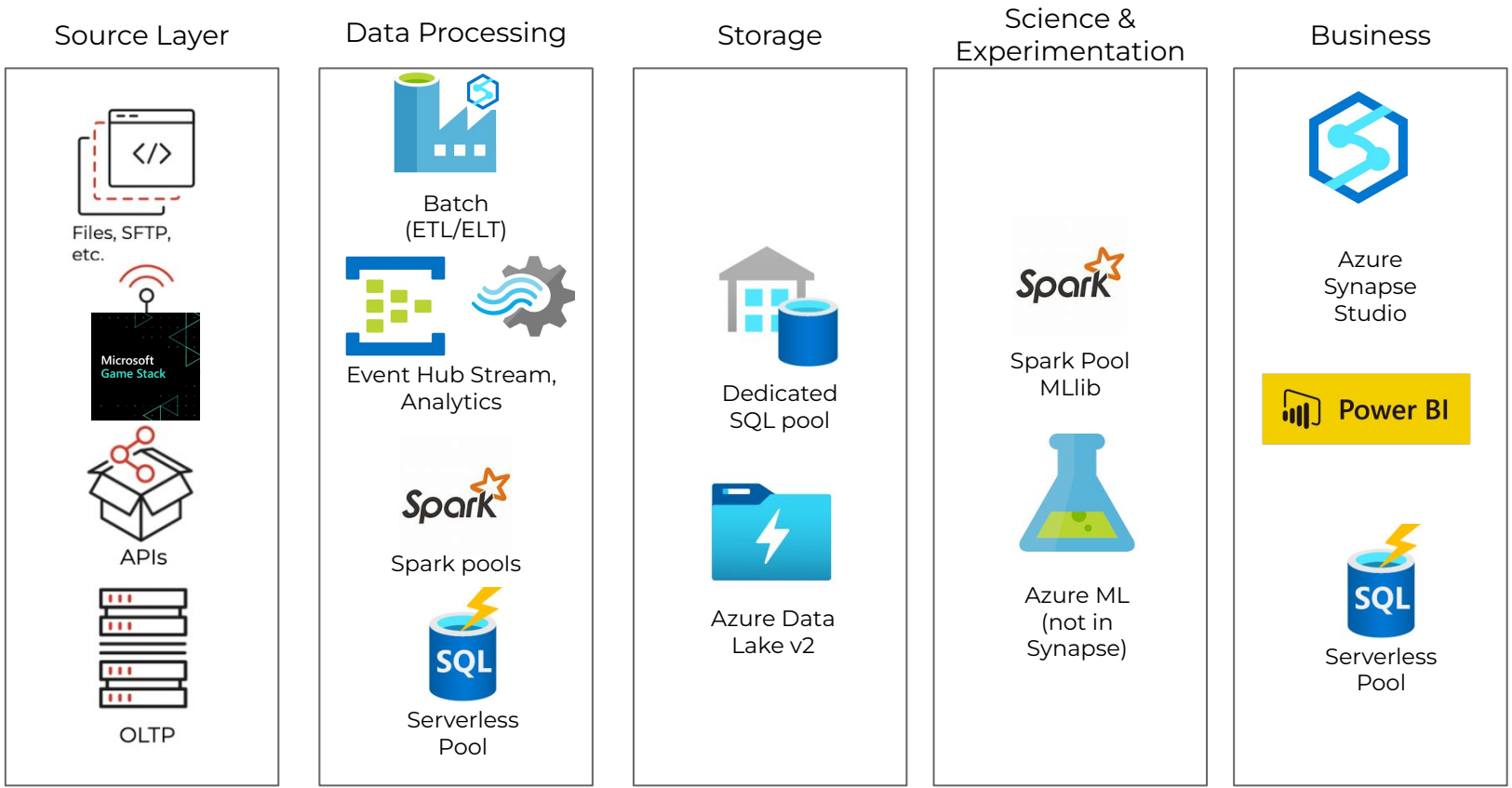
В Microsoft нет DE роли, поэтому все Оки:)

- Azure DevOps
- Visual Studio, Visual Code
- Yaml pipelines

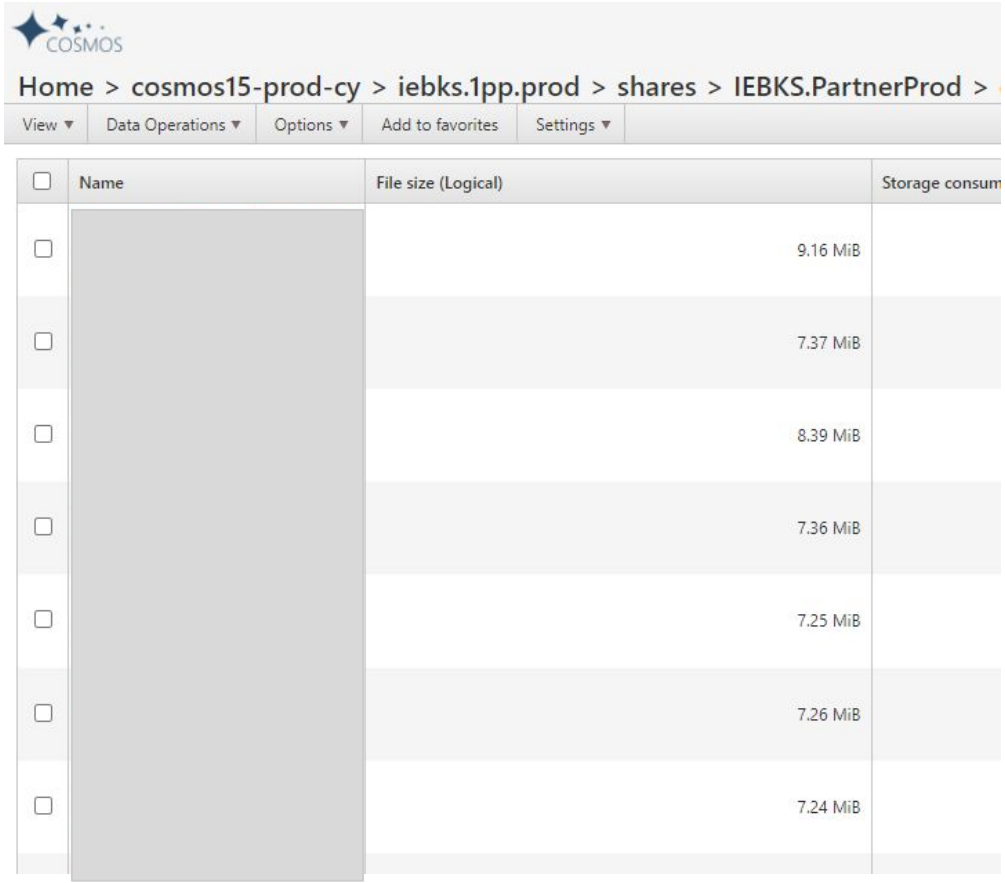
Типичное решение Azure Data Explorer (ADX)



Иногда такое Microsoft Azure Synapse



Microsoft Data Lake

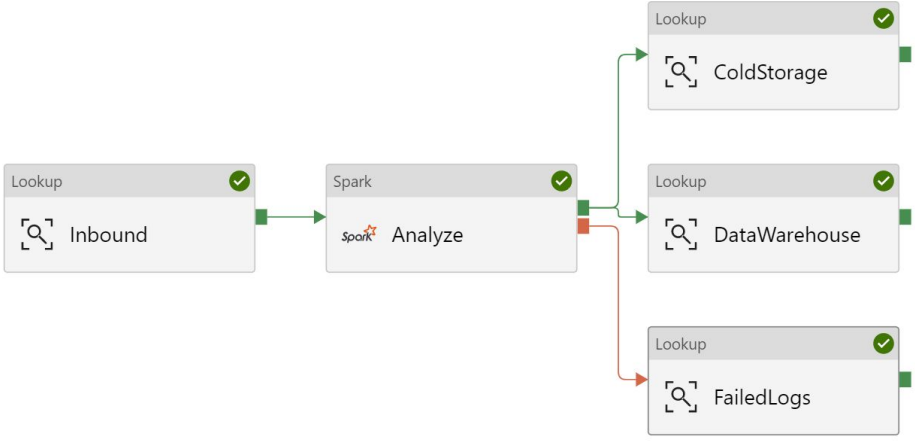


Scope script

```
test_input =
 EXTRACT
   FirstName : string,
   LastName : string,
   Age : int
 FROM
   "/local/Samples/SampleData/test_input.tsv"
   USING DefaultTextExtractor();

OUTPUT test_input
 TO "/local/users/<your alias>/output.tsv"
 USING DefaultTextOutputter();
```

Data Factory Pipeline as a Code (C#)




```
ions.Generic;
e.Management.DataFactory.Models;
aFactory.Shared;

eDataFactory.Halifax

ass HalifaxPipeline : SharedPipeline

c HalifaxPipeline[] AllPipelines => new[]

Cook
new HalifaxPipeline
{
    Name = "pl_Halifax_Cook",
    PipelineResource = new PipelineResource
    {
        Folder = new PipelineFolder("Halifax"),
        Parameters = new Dictionary<string, ParameterSpecification> { { "windowStart", new ParameterSpecificat
        Activities = new List<Activity>
        {
            //Pre-Processing
            new GetMetadataActivity
            {
                Name = "GetMetadata-RtepClient-PreProcessing",
                Dataset = new DatasetReference()
                {
                    ReferenceName = "ds_Halifax_DataLakeStorage_Binary",
                    Parameters = new Dictionary<string, object> {
                        { "filePath", SharedSettings.DateHourFilePath("data-temp/rtep/client") },
                        { "fileName", new Expression ("@coalesce(null)") }
                    }
                },
```

Metadata ETL

Select Queries.sql...ulandrew.com (126))

```
1  SELECT * FROM [procfwk].[ProcessingStageDetails]
2  SELECT * FROM [procfwk].[PipelineProcesses]
3  SELECT * FROM [procfwk].[PipelineParameters]
4
```

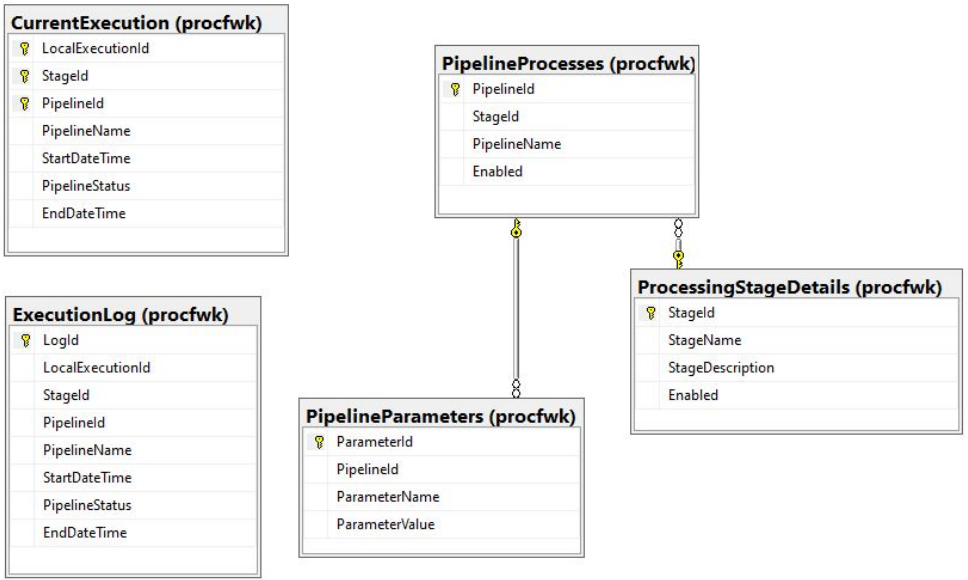
135 %

Results | Messages

	StageId	StageName	StageDescription	Enabled
1	1	Extract	Ingest all data from source systems.	1
2	2	Transform	Transform ingested data and apply business logic.	1
3	3	Load	Load transformed data into semantic layer.	1

	PipelineId	StageId	PipelineName	Enabled
1	1	1	Stage 1-1	1
2	2	1	Stage 1-2	1
3	3	2	Stage 2-1	1
4	5	2	Stage 2-2	1
5	7	2	Stage 2-3	1
6	9	3	Stage 3-1	1

	ParameterId	PipelineId	ParameterName	ParameterValue
1	1	9	TestParam1	Frank
2	2	9	TestParam2	Bob



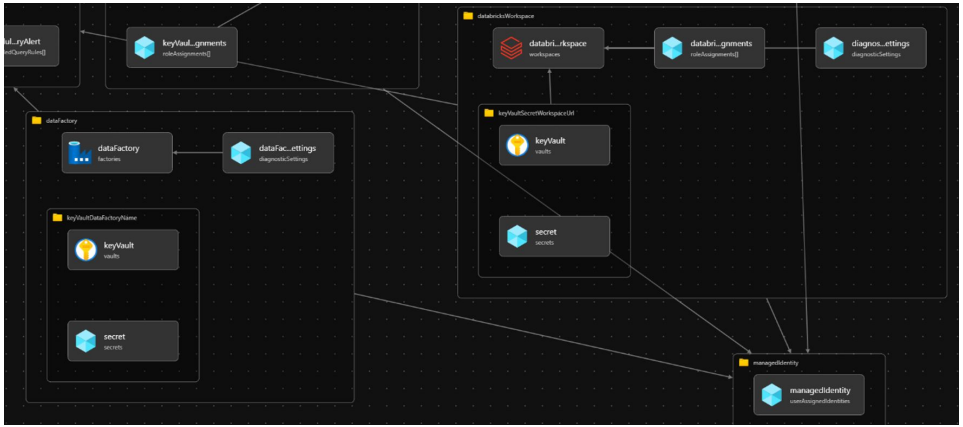
https://mrpaulandrew.com/2020/03/02/creating-a-simple-staged-metadata-driven-processing-framework-for-azure-data-factory-pipelines-part-2-of-4/

Infrastructure as a Code – Azure Bicep

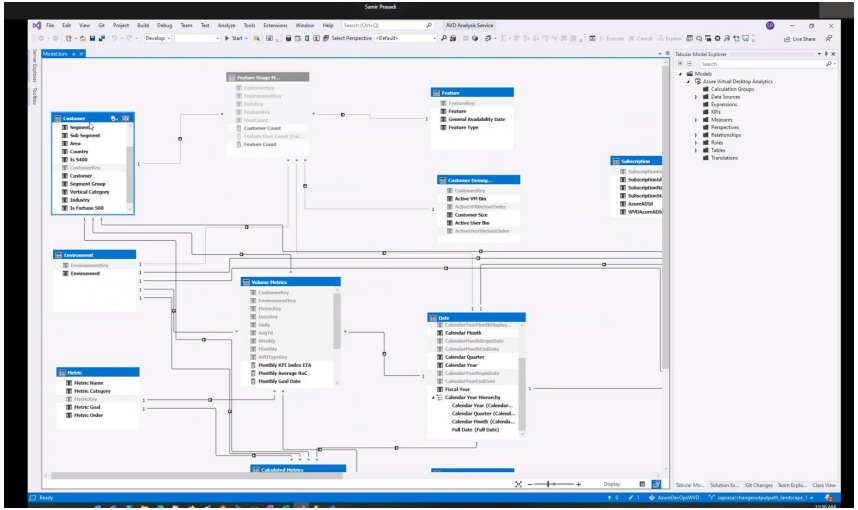
databricks.bicep

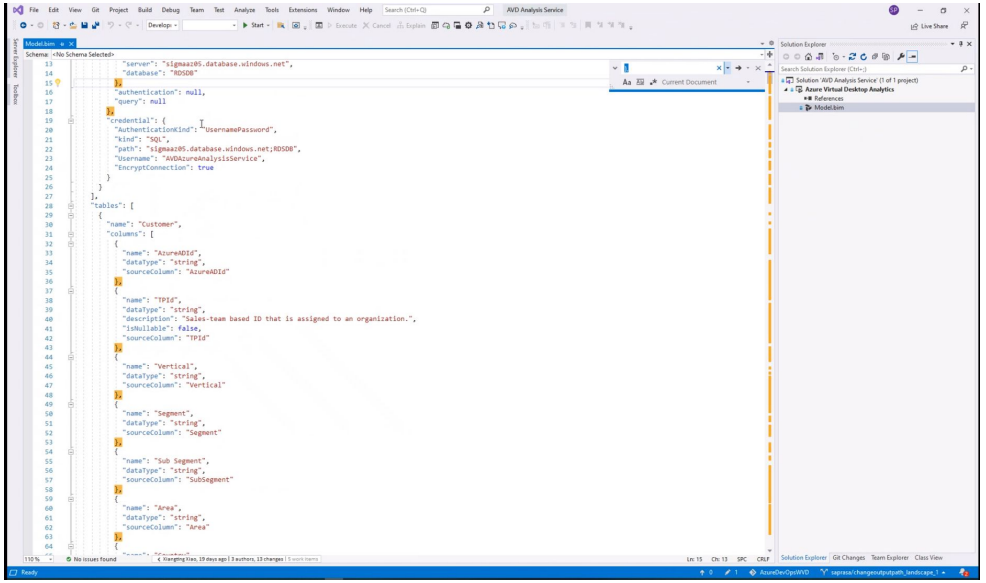
modules > databricks.bicep > ...

```
1  targetScope = 'resourceGroup'
2
3  @description('Location for all resources.')
4  param location string
5
6  @description('The name of the network security group to create.')
7  param workspaceName string
8
9  var managedResourceGroupName = 'databricks-rg-${workspaceName}-${uniqueString(workspaceName, resourceGroup().id)}'
10
11 resource databricksWorkspace 'Microsoft.Databricks/workspaces@2021-04-01-preview' = {
12   name: workspaceName
13   location: location
14   properties: {
15     managedResourceGroupId: subscriptionResourceId('Microsoft.Resources/resourceGroups', managedResourceGroupName)
16   }
17 }
18
19 output workspaceUrl string = databricksWorkspace.properties.workspaceUrl
20
```



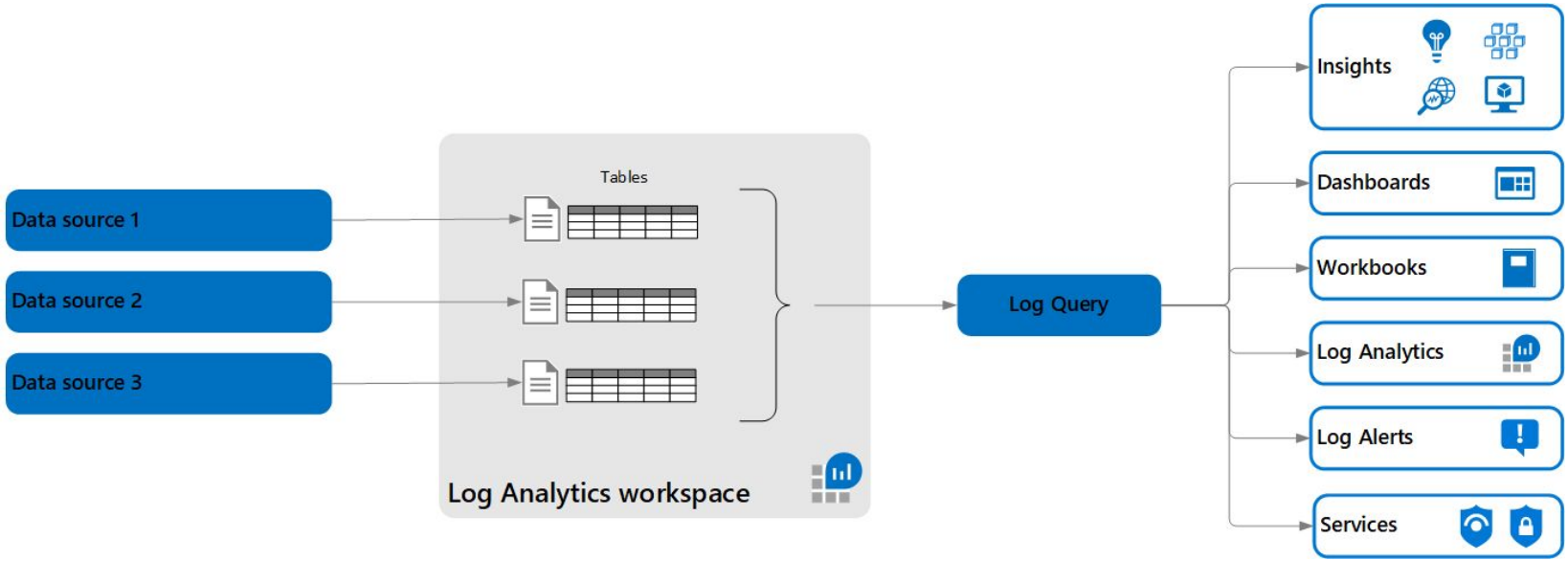
BI as a Code




{
"server": "sigmaaz05.database.windows.net",
"database": "RDSDB"
},
"authentication": null,
"query": null
},
"credential": {
"AuthenticationKind": "UsernamePassword",
"kind": "SQL",
"path": "sigmaaz05.database.windows.net;RDSDB",
"Username": "AVDAzureAnalysisService",
"EncryptConnection": true
}
}
},
"tables": [
{
"name": "Customer",
"columns": [
{
"name": "AzureADId",
"dataType": "string",
"sourceColumn": "AzureADId"
},
{
"name": "TPId",
"dataType": "string",
"description": "Sales-team based ID that is assigned to an organization.",
"isNullable": false,
"sourceColumn": "TPId"
},
{
"name": "Vertical",
"dataType": "string",
"sourceColumn": "Vertical"
},
{
"name": "Segment",
"dataType": "string",
"sourceColumn": "Segment"
},
{
"name": "Sub Segment",
"dataType": "string",
"sourceColumn": "SubSegment"
},
{
"name": "Area",
"dataType": "string",
"sourceColumn": "Area"
},

Alerting

Azure Log Analytics (по сути ADX) умеет собирать данных из всех сервисов Azure и позволяет создавать нам Alerts на ошибки с использование Query. Также можно создавать свои собственные Alerts.



Мои выводы

- У Microsoft своя экосистема
- Нужно любить продукты Microsoft
- Отличная интеграция между продуктами
- Часто приходится работать с сырыми продуктами (DOGFOOD)
- Не самые высокие зарплаты, но хорошие бенефиты
- В Microsoft хорошо попадать “молодым” специалистам в начале карьеры.
- Отличный опыт как делать решения на Azure