



Microsoft Data Platform
Business Intelligence Analytics
Conference

Auckland, New Zealand
18-20 February 2019

www.difinity.co.nz



#Difinity

18th – 20th Feb 2019

<http://difinity.co.nz>



Data wrangling with Azure Databricks

Regan Murphy

Software Engineer, Microsoft

 @nzregs  regan.murphy@microsoft.com

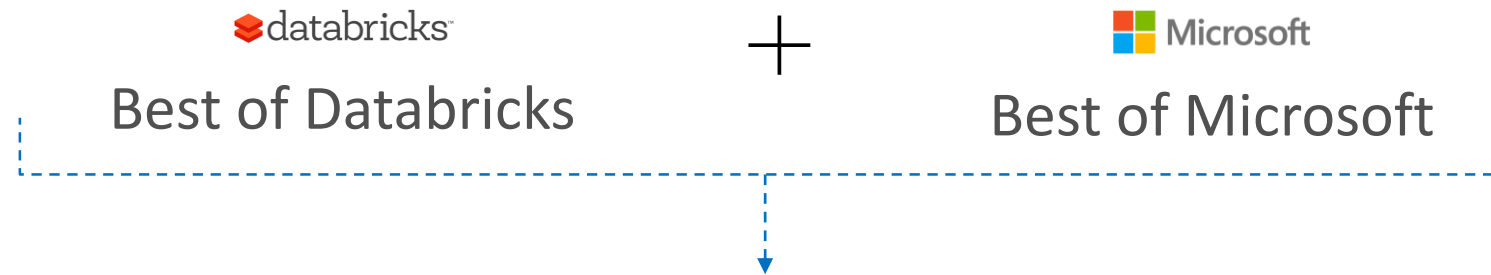
Agenda

- Intro to Azure Databricks
- Scenario
- Technical Demos
 - Secrets
 - Hot Path to Power BI
 - Structured Streaming – Azure Event Hubs Ingestion
 - Warm Path to Cosmos DB
 - Structured Streaming – Azure Event Hubs Ingestion
 - Cold Path, via storage, to SQL Data Warehouse
 - Azure Event Hub Capture, Blob Storage, ADLS Gen 2, Azure Data Factory


Intro to Azure Databricks

What is Azure Databricks ?


A fast, easy and collaborative Apache® Spark™ based analytics platform optimized for Azure



 Designed in collaboration with the founders of Apache Spark

 One-click set up; streamlined workflows

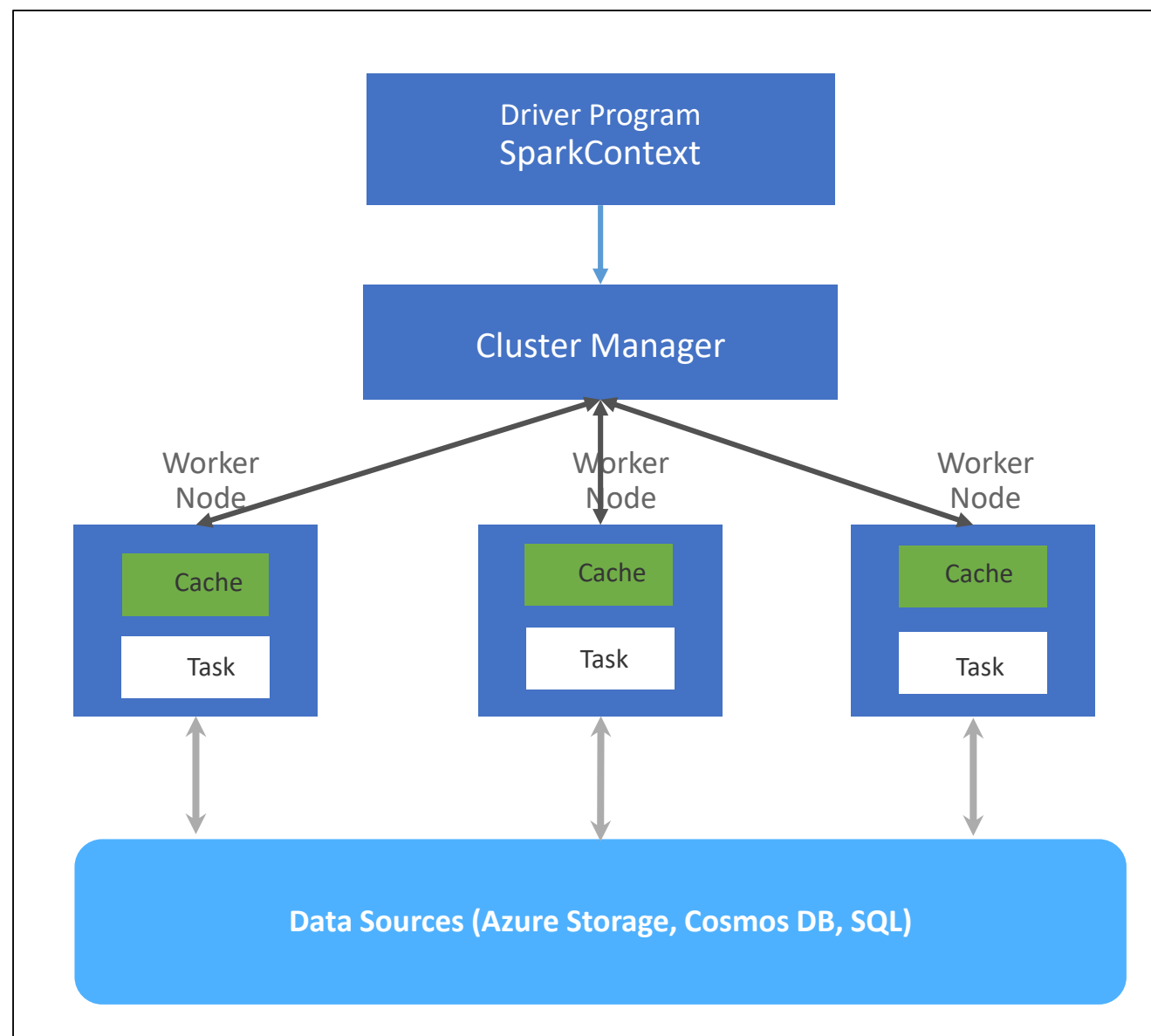
 Interactive workspace that enables collaboration between data scientists, data engineers, and business analysts.

 Native integration with Azure services (Power BI, SQL DW, Cosmos DB, Blob Storage, ADF, SQL DB, AAD)

 Enterprise-grade Azure security (Active Directory integration, compliance, enterprise-grade SLAs – 99.95%)

Spark Architecture & Dataframes

- 'Driver' runs the user's 'main' function and executes the various parallel operations on the worker nodes.
- To take advantage of Spark – you use Dataframes as the data structure.
- Once your Data is in the DataFrame – Spark can parallelize operations.
- The Dataframes support both batch and streaming data.
- The results of the operations are collected by the driver

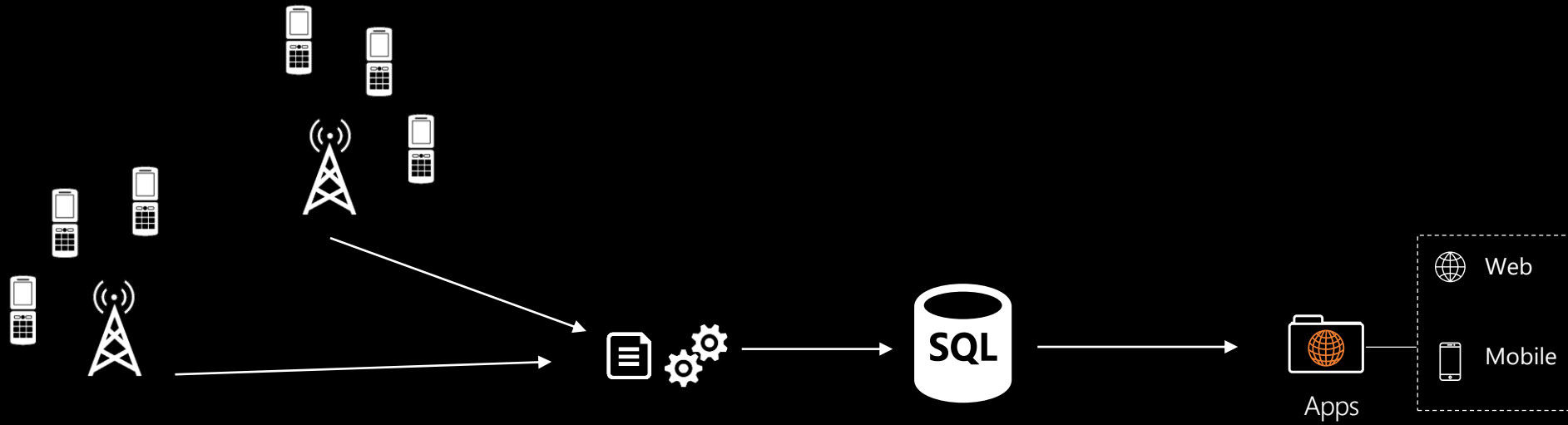


Databricks Notebooks 101

Some basics

The Scenario

Mobile Telco Billing



Check your usage

Unbilled (Current)

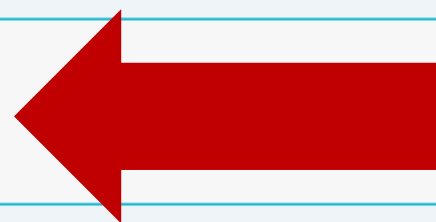
Latest bill

Previous bill

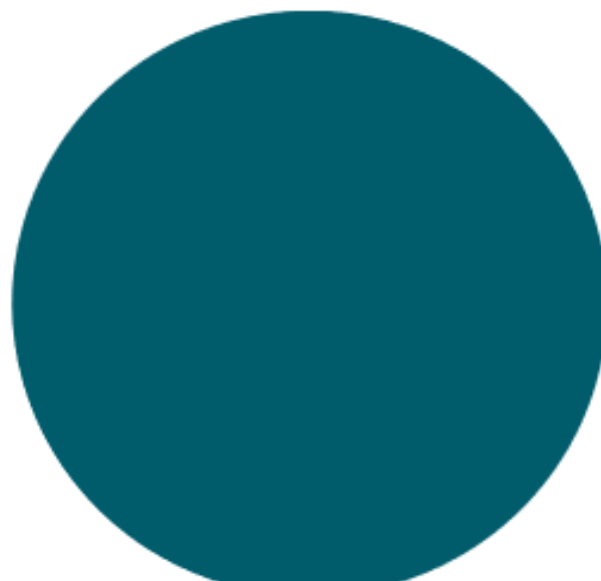


Just so you know

Usage information for the last few hours may not be displayed.



Usage by spend



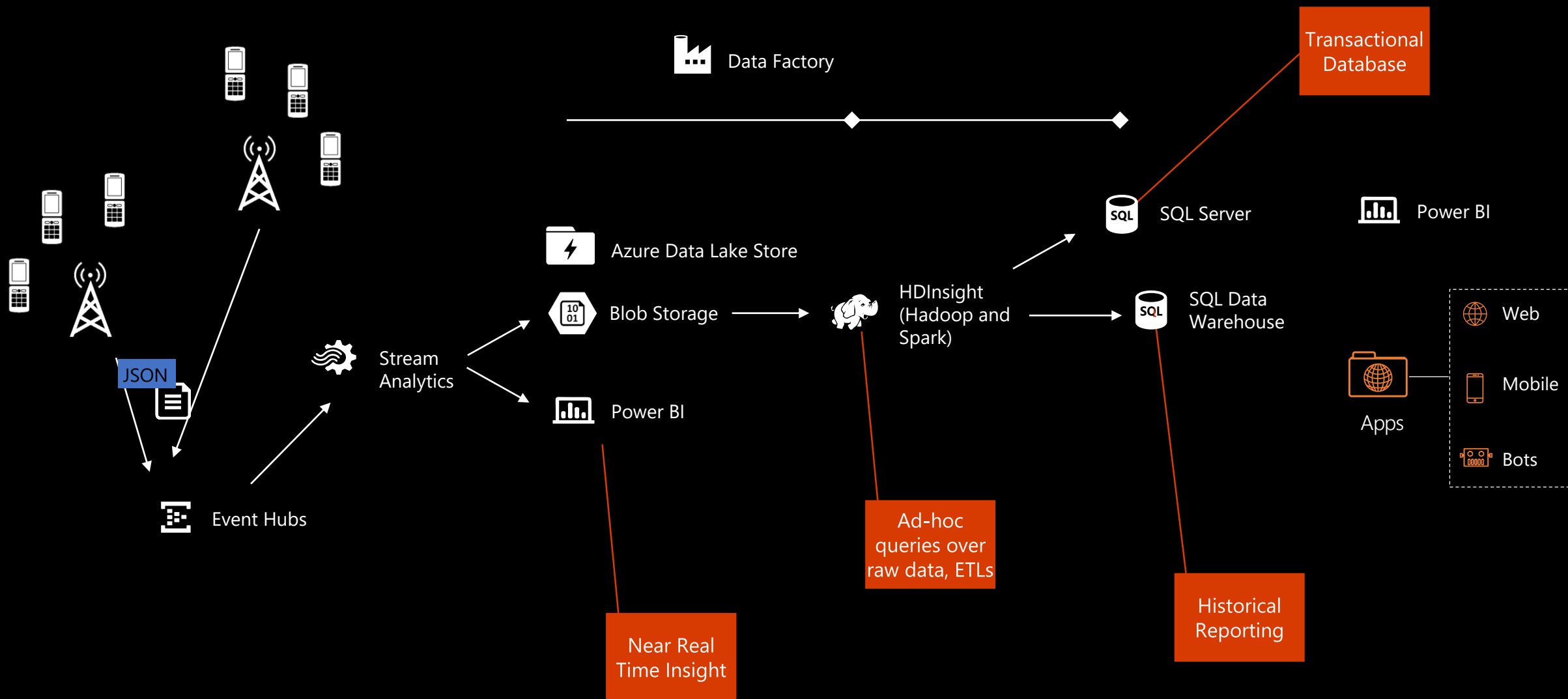
All prices excl. GST

International TXTs	2	\$0.34
Data	3703.51 MB	\$0.00
Calls	19:00 MINs	\$0.00
TXTs	14	\$0.00

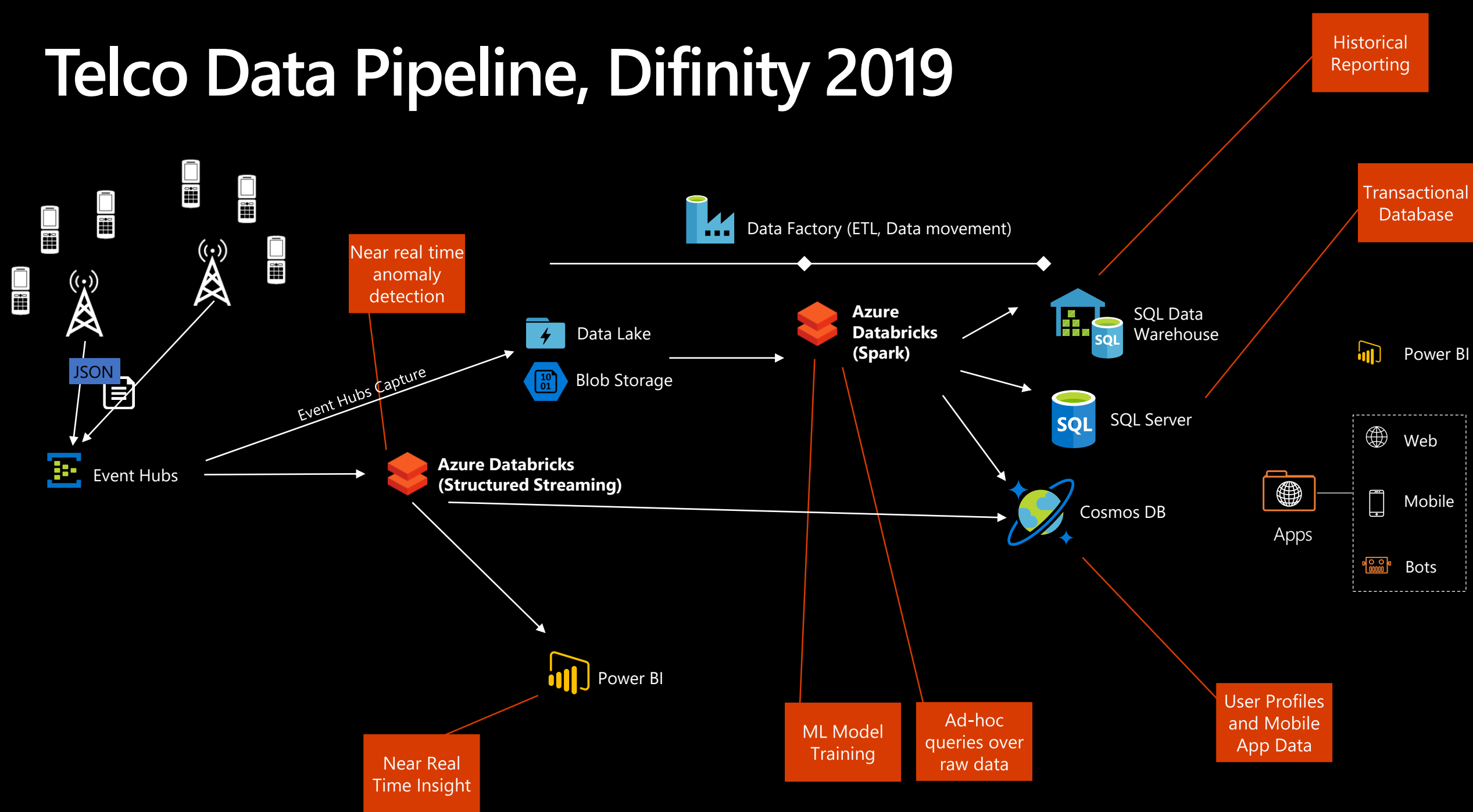
Total out of plan spend

\$0.34

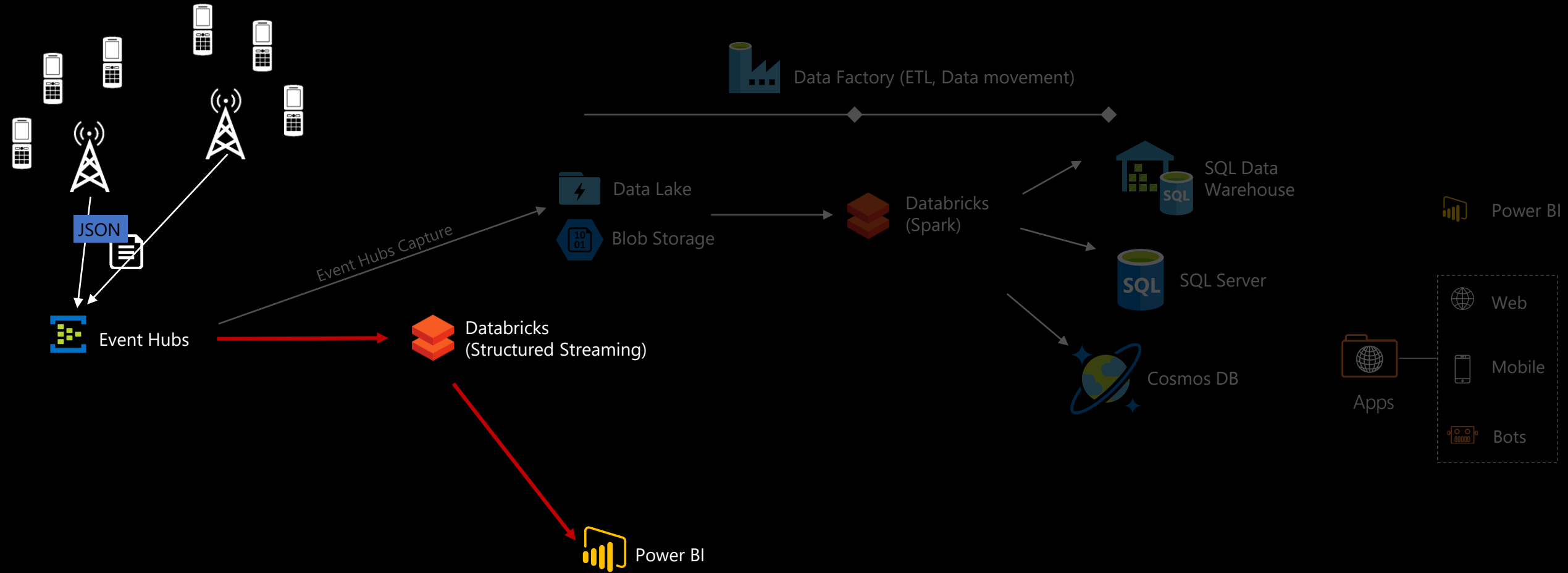
Telco Data Pipeline, Dfinity 2018



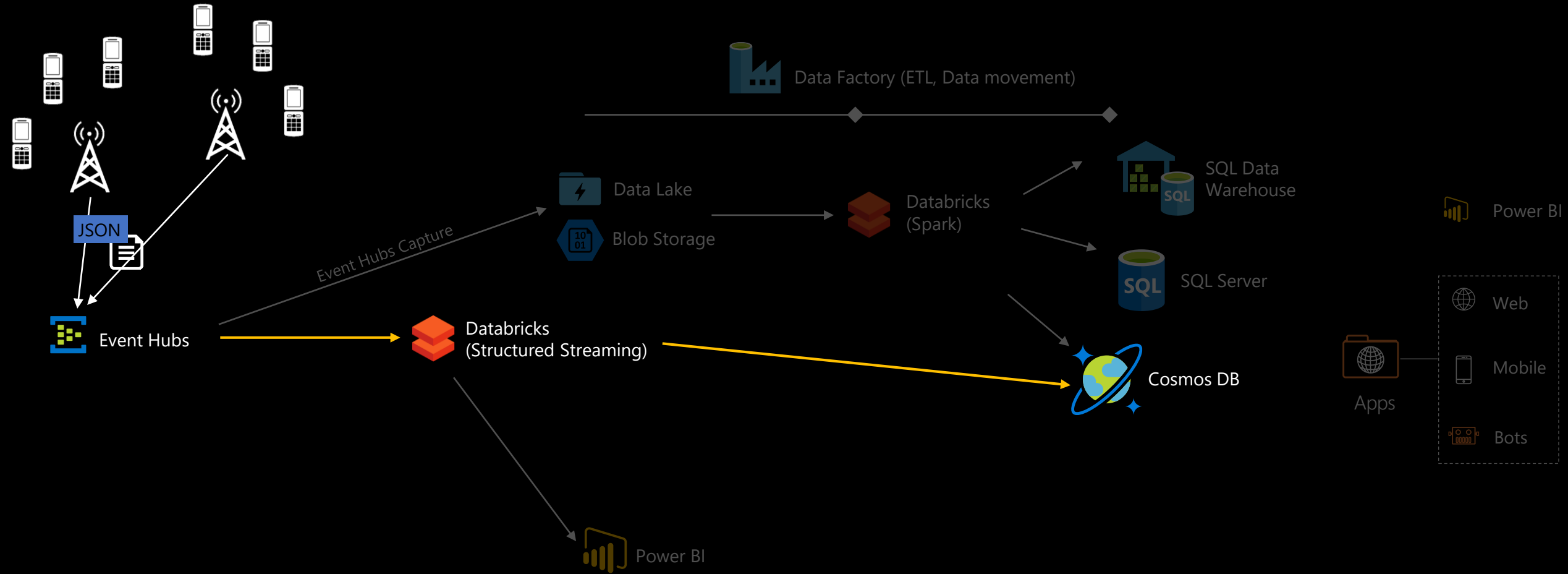
Telco Data Pipeline, Dfinity 2019



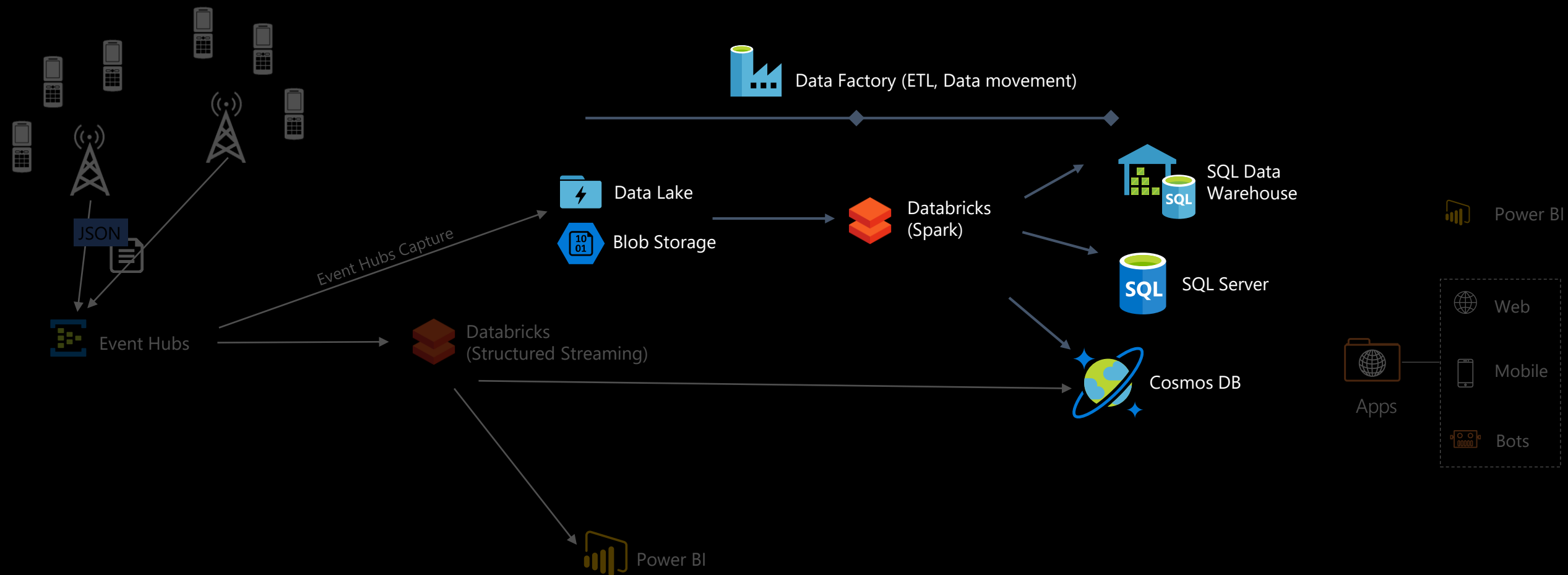
Lambda – the **hot** path



Lambda – the **warm** path



Lambda – the cold path



Building the solution

Secrets

Don't leak your credentials

Secret Management

When connecting to services from within Azure Databricks, you will frequently need to pass credentials. Best practice would be to keep these separate from your code.

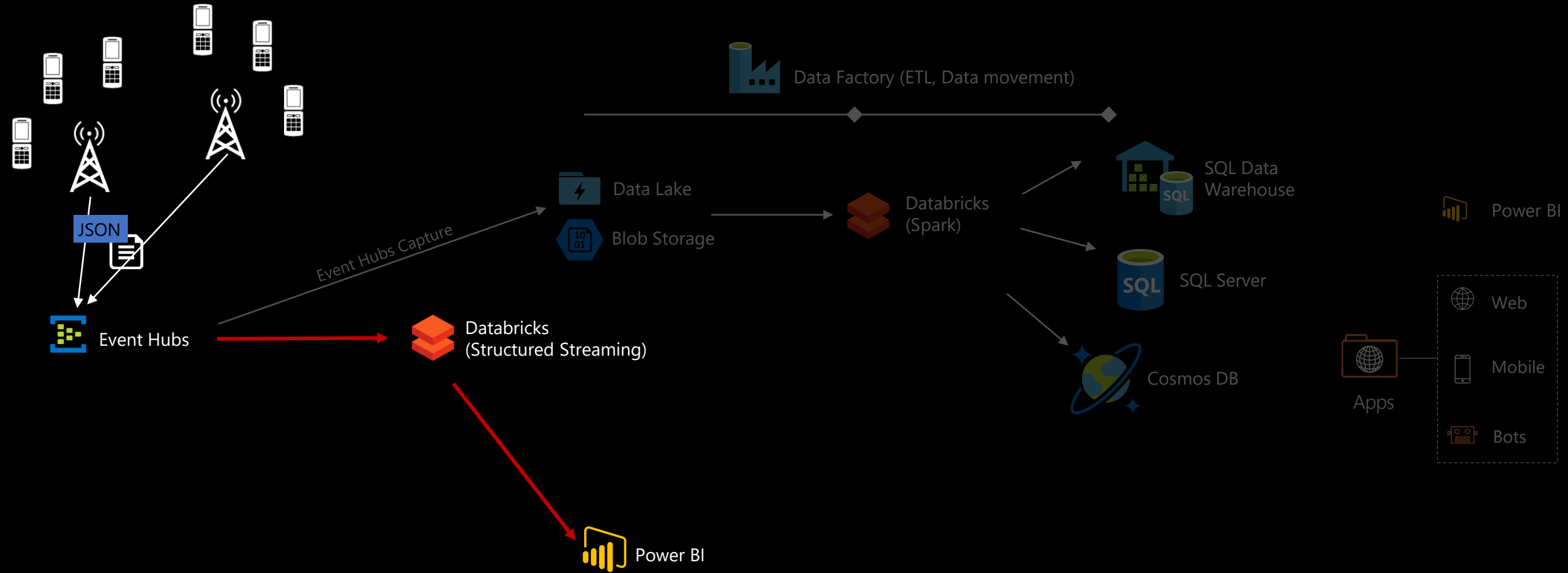
Options (least secure ☹️ to most secure 😊):

- Plain text in notebooks << *please avoid this!*
- Databricks Secret Scopes
- Azure Key Vault backed Secret Scopes

Hot Path

Azure Event Hubs --> Power BI

Lambda – the **hot** path



Hot Path

- Azure Event Hubs Connector for Apache Spark
 - Connect to the Azure Event Hub and ingest real-time events
 - <https://github.com/Azure/azure-event-hubs-spark>
- Spark Scala & Spark SQL
 - Enrich streaming dataset using reference datasets
 - Aggregate using SQL Window functions, if necessary
- Custom ForeachWriter
 - Write each row of the enriched, aggregated, dataset out to a Power BI custom streaming dataset using HTTP Post to the Power BI custom API endpoint

Power BI Custom Streaming Dataset API

```
// This is called for each row after open() has been called.
// This implementation sends one row at a time.
// A more efficient implementation can be to send batches of rows at a time.
//
def process(row: Row) = {
  val rowAsMap = row.getValuesMap(row.schema.fieldNames)
  val dummyrow = "[{"eventDate\" : \"\"+LocalDateTime.now().toString(), \"stringData\" : \"\"+LocalDateTime.now().toString()}]"

  val rowToSend = "[{"towerId\" : \"\"+row.getAs(\"towerId\").toString()+\", \"eventDate\" : \"\"+row.getAs(\"eventDate\").toString()+\", \"towername\" : \"\"+row.getAs(\"Name\").toString()+\", \"toweraddress\" : \"\"+row.getAs(\"Address\").toString()+\", \"towercity\" : \"\"+row.getAs(\"City\").toString()+\", \"towerlongitude\" : \"\"+row.getAs(\"Longitude\").toString()+\", \"towerlatitude\" : \"\"+row.getAs(\"Latitude\").toString()+\", \"imei\" : \"\"+row.getAs(\"imei\").toString()+\", \"fromNumber\" : \"\"+row.getAs(\"fromNumber\").toString()+\", \"toNumber\" : \"\"+row.getAs(\"toNumber\").toString()+\", \"billingType\" : \"\"+row.getAs(\"billingType\").toString()+\", \"duration\" : \"\"+row.getAs(\"duration\").toString()+\", \"bytes\" : \"\"+row.getAs(\"bytes\").toString()+\", \"protocol\" : \"\"+row.getAs(\"protocol\").toString()+\", \"port\" : \"\"+row.getAs(\"port\").toString()+\", \"uri\" : \"\"+row.getAs(\"uri\").toString()+\", \"cost\" : \"\"+row.getAs(\"cost\").toString()+\", \"}]]"

  // "+row.getAs(\"fromNumber\").toString()+\"

  println("row to send: " + rowToSend)

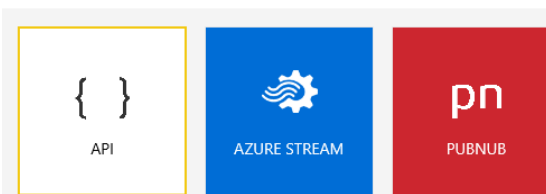
  val post = new HttpPost(PowerBIEndpoint_DifinityTelecomTowerFeedLive)
  post.setHeader("Content-type", "application/json")
  post.setEntity(new StringEntity(rowToSend))

  val client: CloseableHttpClient = HttpClientBuilder.create().build();

  val response: HttpResponse = client.execute(post)
  println(response.getStatusLine.getStatusCode, response.getStatusLine.getReasonPhrase)
}
```

New streaming dataset

Choose the source of your data



Edit streaming dataset

Create a streaming dataset and integrate our API into your device or application to send data. [Learn more about the API.](#)

* Required

Dataset name *

Values from stream *

towerId	Text	✕
eventDate	DateTime	✕
imei	Text	✕
fromNumber	Text	✕
toNumber	Text	✕

API info on DifinityTelecomTowerFeed...

Use the API endpoint URL and one of the examples shown below to send data to your streaming dataset. For more information, [read our API documentation and integration guide.](#)

Push URL

Raw

cURL

PowerShell

```
[
  {
    "towerId" : "AAAAA55555",
    "eventDate" : "2019-02-19T10:38:46.567Z",
    "imei" : "AAAAA55555",
    "fromNumber" : "AAAAA55555",
    "toNumber" : "AAAAA55555",
    "billingType" : "AAAAA55555",
    "duration" : 98.6,
    "bytes" : 98.6,
    "protocol" : 98.6,
    "port" : 98.6,
    "uri" : "AAAAA55555",
    "cost" : 98.6,
    "towername" : "AAAAA55555",
    "toweraddress" : "AAAAA55555",
    "towerlongitude" : "AAAAA55555",
    "towerlatitude" : "AAAAA55555",
    "towercity" : "AAAAA55555"
  }
]
```

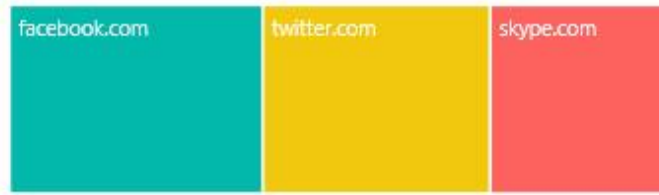
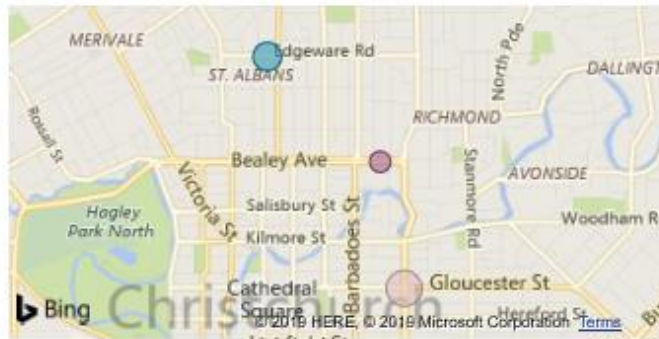
Auckland



Wellington



Christchurch



01_DifinityTelecom_HotPath (Scala)

Attached: ● DifinityInteractiveCl... File View: Code Permissions Stop Execution Clear

Schedule Comments Revision history

display(df)

[Cancel](#)

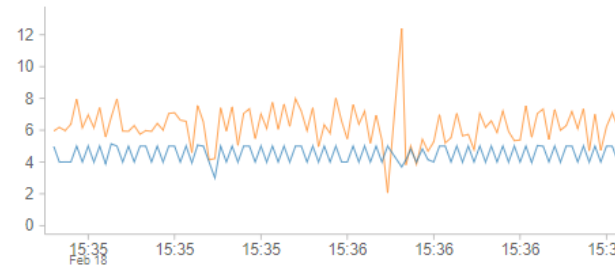
▶ (1) Spark Jobs

▼ ● display_query_1 (id: 6bf4b742-d33b-4d35-9229-ee0633fe8d0d) *Last updated: About now*

[Dashboard](#) [Raw Data](#)

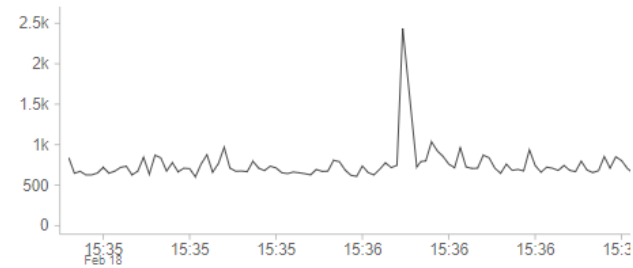
Input vs. Processing Rate
records per second

5 rec/s 7.8 rec/s
Input rate Processing rate



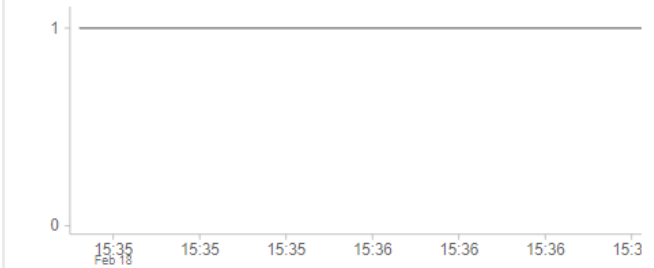
Batch Duration
in milliseconds

741.9 ms 642 ms
Average Latest



Aggregation State

1
Distinct keys



towerId	eventDate	imei	fromNumber	toNumber	billingType	duration	bytes	protocol	port	uri	cost
100000	2019-02-18T02:27:49.000+0000		215550100	null	data	0	403	6	80	twitter.com	0
100004	2019-02-18T02:27:50.000+0000		215553537	215558319	call	56	0	0	0	null	0
239042	2019-02-18T02:27:50.000+0000		215559726	215558302	sms	0	0	0	0	null	0
100004	2019-02-18T02:27:50.000+0000		215555845	215557244	call	14	0	0	0	null	0
209653	2019-02-18T02:27:51.000+0000		215559641	null	data	0	257	6	80	facebook.com	0
209653	2019-02-18T02:27:51.000+0000		215554028	null	data	0	178	6	80	twitter.com	0
244121	2019-02-18T02:27:52.000+0000		215550782	215556546	call	43	0	0	0	null	0
100006	2019-02-18T02:27:52.000+0000		215559004	null	data	0	949	6	80	twitter.com	0
244107	2019-02-18T02:27:52.000+0000		215551547	215551555	sms	0	0	0	0	null	0

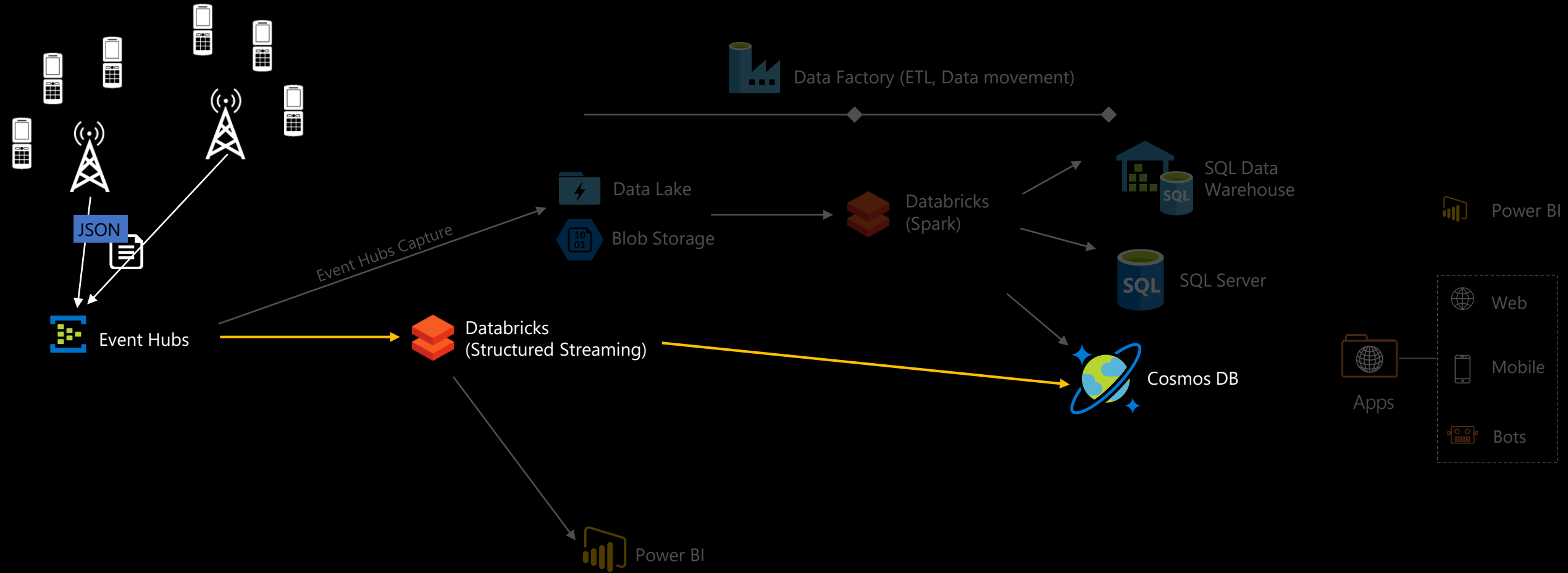
Showing the first 1000 rows.

Table Chart

Warm Path

Azure Event Hubs --> Cosmos DB

Lambda – the **warm** path



Warm Path

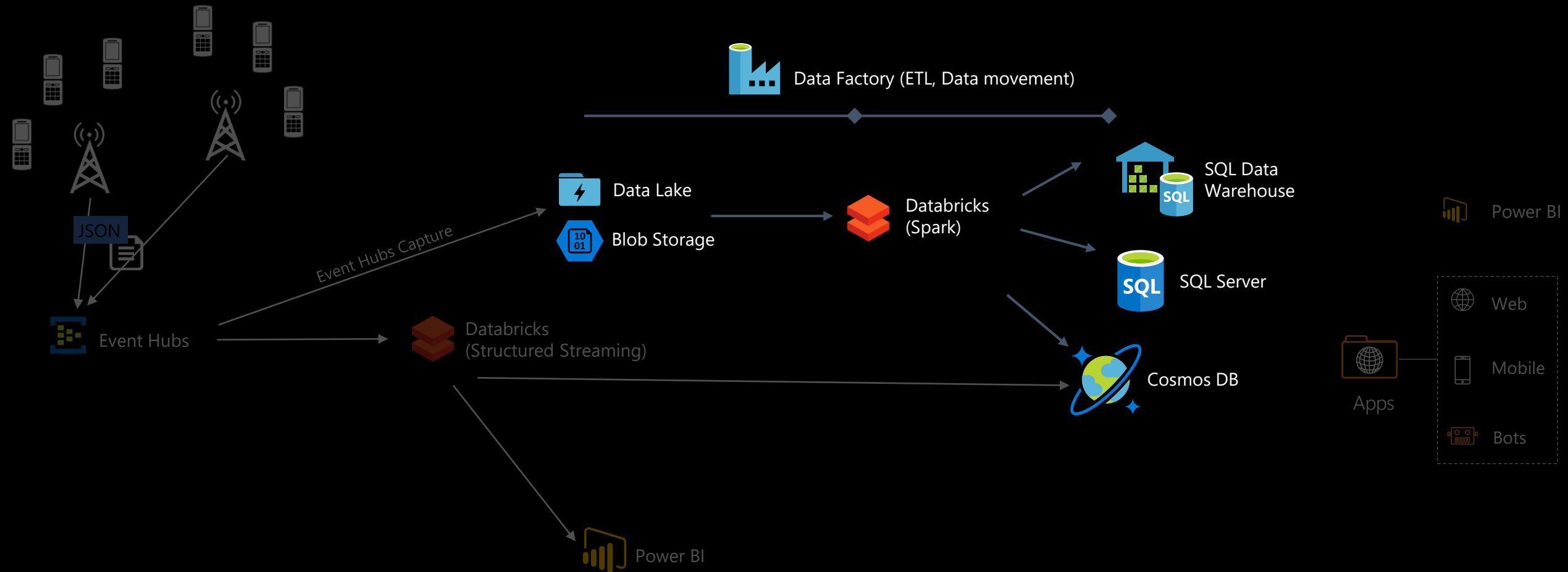
- Azure Event Hubs Connector for Apache Spark
 - Connect to the Azure Event Hub and ingest real-time events
 - <https://github.com/Azure/azure-event-hubs-spark>
- Spark Scala & Spark SQL
 - Enrich streaming dataset using reference datasets
 - Aggregate using SQL Window functions
- Azure Cosmos DB Connector for Apache Spark
 - Write enriched, aggregated, streaming dataset out to Cosmos DB
 - <https://github.com/Azure/azure-cosmosdb-spark>

Cold Path

Azure Event Hub Capture

- > AVRO files on Azure Blob Storage
- > Azure Datalake Gen2
- > SQL Data Warehouse

Lambda – the cold path



Cold Path

- Azure Event Hubs Capture
 - Built-in option to capture to Azure Blob Storage in AVRO format
 - <https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-capture-overview>

ORCHESTRATED USING AZURE DATA FACTORY:

- Spark Scala & Spark SQL
 - Load AVRO files into dataframe
 - Enrich and Aggregate using SQL
 - Write PARQUET files out to storage (blob or datalake)
- Azure SQL Data Warehouse
 - Load the data from the PARQUET files into SQL Data Warehouse

Microsoft Ignite Sessions on YouTube™

- An Introduction to big data processing with Azure Databricks
 - THR2182 - Nishant Thacker - <http://bit.ly/ignite-thr2182>
- Azure Databricks for data engineers and data developers
 - BRK3313 – Bhanu Prakash - <http://bit.ly/ignite-brk3313>
- Real-time analytics with Azure Databricks and Azure Event Hubs
 - BRK3202 – Bhanu Prakash and Abhinav Garg - <http://bit.ly/ignite-brk3202>



Regan Murphy

Software Engineer, Microsoft

 @nzregs  regan.murphy@microsoft.com

<https://github.com/nzregs/DifinityTelecom>

<https://hub.docker.com/r/nzregs/datagenntecore>

<https://github.com/nzregs/slidedecks>

Evaluate Sessions and Win a Prize!



<https://www.surveymonkey.com/r/5LR9LFB>

Thanks to our sponsors

Platinum Sponsors



Silver



Exhibitor

