

Springboard Open-ended Capstone Overview

This project starts with finding a large data set, then builds a data pipeline to explore and extract valuable information from a small sample data set on a local Apache Spark system, then scale up and deploy it on to Microsoft Azure Databricks for production.

Short Term Rental Facts

- Short term rentals have exploded with the help of platforms such as [airbnb.com](https://www.airbnb.com), [booking.com](https://www.booking.com), [vrbo.com](https://www.vrbo.com).
- Both travellers and hosts need information to find or host a space.
- This project collects large historical, current and future data to extract valuable information.

Data Pipeline Design

- The data pipeline should be run quarterly based on the chosen data set.
- Data acquisition: python program that downloads the data set from a web source.
- Data ingestion: the data set is read, cleaned, then written to storage in partitions.
- Data analytics: Focused on the reviews of each property. It extracts the most frequent used words from the reviews to give people some idea without reading the actual reviews.

Azure Databricks Spark Cluster

Microsoft Azure | Databricks

Create a cluster

You'll use compute resources (clusters) to run your commands.

Click 'Create cluster' and use our [best practices guide](#) to set up your cluster.

Clusters / New Compute

New Cluster

Cancel

Create Cluster

DBU / hour: 2.25 - 6.75

2-8 Workers: 28-112 GB Memory, 8-32 Cores
1 Driver: 14 GB Memory, 4 Cores

Cluster name

strfactsSparkCluster

Cluster mode

Standard

Databricks runtime version

Runtime: 10.4 LTS (Scala 2.12, Spark 3.2.1)

Promotional discount applied to Photon during preview

☐ Use Photon Acceleration Preview

Autopilot options

☒ Enable autoscaling

☒ Terminate after 060 minutes of inactivity

Worker type

Standard_DS3_v2 14 GB Memory, 4 Cores

Min workers 2

Max workers 8

☒ Spot instances

Driver type

Same as worker 14 GB Memory, 4 Cores

DBU / hour: 2.25 - 6.75

Standard_DS3_v2

Advanced options

Azure Blob Storage

Microsoft Azure

Search resources, services, and doc

[Home](#) > [Create a resource](#) >

Create a storage account ...

[Basics](#) [Advanced](#) [Networking](#) [Data protection](#) [Encryption](#) [Tags](#) [Review + create](#)

Project details

Select the subscription in which to create the new storage account. Choose a new or existing resource group to organize and manage your storage account together with other resources.

Subscription *

Azure subscription 1

Resource group *

(New) strfactsRG

[Create new](#)

Instance details

If you need to create a legacy storage account type, please click [here](#).

Storage account name ⓘ *

strfactsblob

Region ⓘ *

(US) West US

Performance ⓘ *

☒ **Standard:** Recommended for most scenarios (general-purpose v2 account)

☐ **Premium:** Recommended for scenarios that require low latency.

Redundancy ⓘ *

Geo-redundant storage (GRS)

☒ Make read access to data available in the event of regional unavailability.

[Review + create](#)

[< Previous](#)

[Next: Advanced >](#)

Blob After Data Acquisition

The screenshot displays the Microsoft Azure Storage browser interface. The top navigation bar includes the Microsoft Azure logo, a search bar, and user information (sxu2012@gmail.com). The main header shows the storage account 'strfactsblob' and the 'Storage browser (preview)' view. The left sidebar contains a navigation menu with categories: Overview, Activity log, Tags, Diagnose and solve problems, Access Control (IAM), Data migration, Events, Storage browser (preview), Data storage (Containers, File shares, Queues, Tables), Security + networking (Networking, Azure CDN, Access keys, Shared access signature, Encryption, Microsoft Defender for Cloud), and Data management (Geo-replication, Data protection, Object replication).

The main content area shows the 'bcontainer1' container. The breadcrumb path is 'Blob containers > bcontainer1 > indata > 2022-03'. The authentication method is 'Access key (Switch to Azure AD User Account)'. A search bar for blobs by prefix is present, along with a filter for 'Only show active blobs'. The table below lists 20 blobs, all of which are 'Block blob' type and 'Hot (Inferred)' access tier. The columns are Name, Last modified, Access tier, Blob type, Size, Lease state, and a more options menu.

Name	Last modified	Access tier	Blob type	Size	Lease state
☐ washington-dc-neighbourhoods.csv	6/9/2022, 3:50:09 PM	Hot (Inferred)	Block blob	1.93 KiB	Available
☐ washington-dc-neighbourhoods.geojson	6/9/2022, 3:50:09 PM	Hot (Inferred)	Block blob	356.34 KiB	Available
☐ washington-dc-reviews.csv	6/9/2022, 3:50:09 PM	Hot (Inferred)	Block blob	5.89 MiB	Available
☐ washington-dc-reviews.csv.gz	6/9/2022, 3:50:07 PM	Hot (Inferred)	Block blob	32.31 MiB	Available
☐ western-australia-calendar.csv.gz	6/9/2022, 3:50:12 PM	Hot (Inferred)	Block blob	9.55 MiB	Available
☐ western-australia-listings.csv	6/9/2022, 3:50:16 PM	Hot (Inferred)	Block blob	1.3 MiB	Available
☐ western-australia-listings.csv.gz	6/9/2022, 3:50:11 PM	Hot (Inferred)	Block blob	6.68 MiB	Available
☐ western-australia-neighbourhoods.csv	6/9/2022, 3:50:18 PM	Hot (Inferred)	Block blob	1.56 KiB	Available
☐ western-australia-neighbourhoods.geojson	6/9/2022, 3:50:19 PM	Hot (Inferred)	Block blob	7.14 MiB	Available
☐ western-australia-reviews.csv	6/9/2022, 3:50:18 PM	Hot (Inferred)	Block blob	8.8 MiB	Available
☐ western-australia-reviews.csv.gz	6/9/2022, 3:50:16 PM	Hot (Inferred)	Block blob	39.09 MiB	Available
☐ zurich-calendar.csv.gz	6/9/2022, 3:50:26 PM	Hot (Inferred)	Block blob	1.75 MiB	Available
☐ zurich-listings.csv	6/9/2022, 3:50:28 PM	Hot (Inferred)	Block blob	274.56 KiB	Available
☐ zurich-listings.csv.gz	6/9/2022, 3:50:25 PM	Hot (Inferred)	Block blob	1.02 MiB	Available
☐ zurich-neighbourhoods.csv	6/9/2022, 3:50:29 PM	Hot (Inferred)	Block blob	652 B	Available
☐ zurich-neighbourhoods.geojson	6/9/2022, 3:50:29 PM	Hot (Inferred)	Block blob	509.25 KiB	Available
☐ zurich-reviews.csv	6/9/2022, 3:50:28 PM	Hot (Inferred)	Block blob	989.53 KiB	Available
☐ zurich-reviews.csv.gz	6/9/2022, 3:50:27 PM	Hot (Inferred)	Block blob	5.3 MiB	Available

Blob After Data Ingestion

The screenshot displays the Microsoft Azure Storage browser interface. The top navigation bar shows the Microsoft Azure logo and a search bar. The breadcrumb path is "Home > strfactsblob". The main heading is "strfactsblob | Storage browser (preview)".

The left sidebar contains a search bar and a list of navigation options: Overview, Activity log, Tags, Diagnose and solve problems, Access Control (IAM), Data migration, Events, Storage browser (preview) (selected), Data storage (Containers, File shares, Queues, Tables), Security + networking (Networking, Azure CDN, Access keys, Shared access signature, Encryption, Microsoft Defender for Cloud), and Data management (Geo-replication, Data protection, Object replication).

The main content area shows the "strfactsblob" storage account. The "Blob containers" section is expanded, showing "bcontainer1" selected. The "View all" link is visible. The "Authentication method" is "Access key (Switch to Azure AD User Account)". The search bar is "Search blobs by prefix (case-sensitive)".

The table displays the first 100 items, showing the following columns: Name, Last modified, Access tier, and Blob type. The items listed are:

Name	Last modified	Access tier	Blob type
[...]			
location=NONE			
location=antwerp			
location=asheville			
location=athens			
location=austin			
location=bangkok			
location=barcelona			
location=barossa-valley			
location=barwon-south-west-vic			
location=beijing			
location=belize			
location=bergamo			
location=berlin			
location=bologna			
location=bordeaux			
location=boston			
location=bristol			

Blob After Data Analytics

The screenshot displays the Microsoft Azure Storage browser interface. The top navigation bar shows the Microsoft Azure logo and a search bar. The breadcrumb path is "Home > strfactsblob". The main heading is "strfactsblob | Storage browser (preview)".

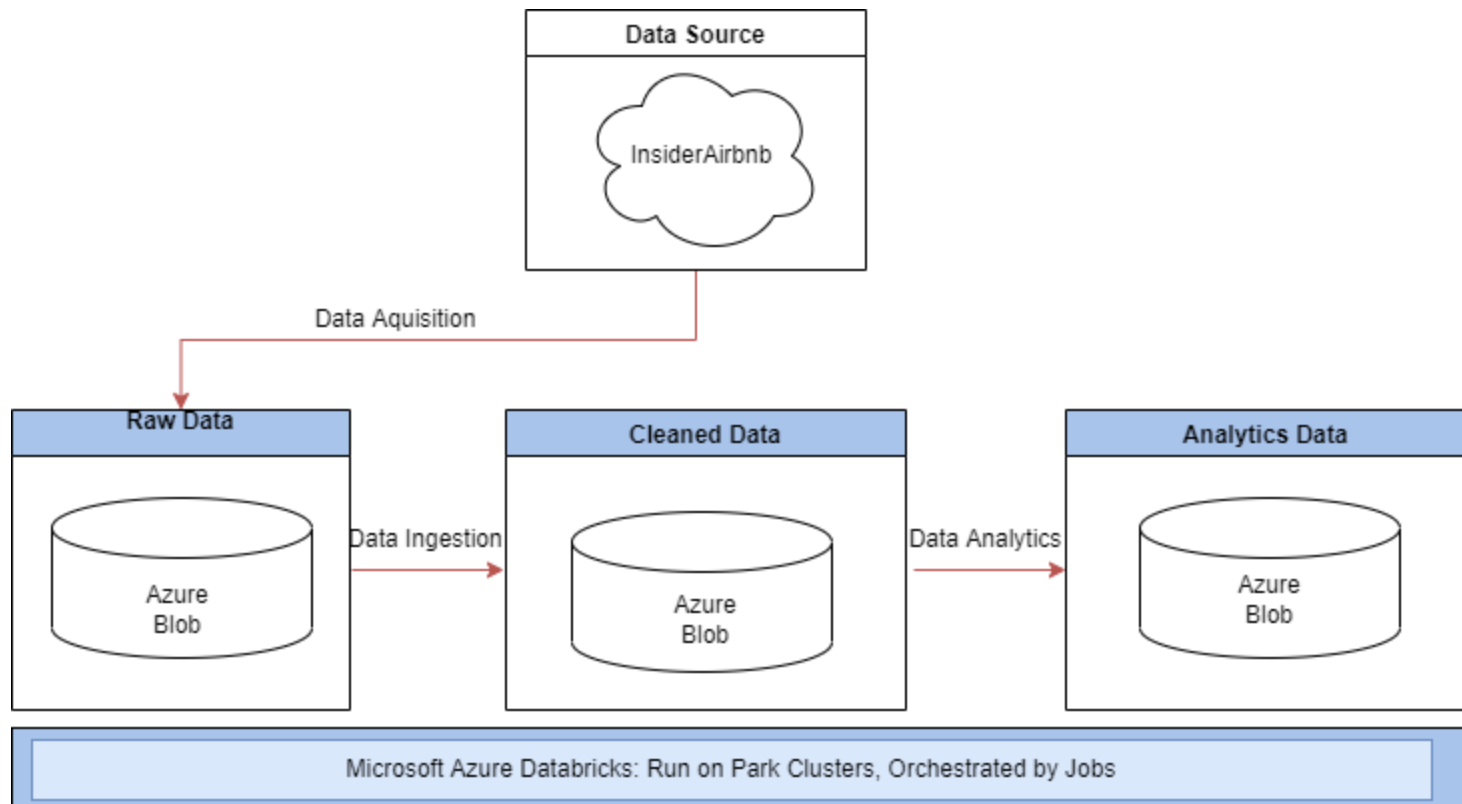
The left sidebar contains a navigation menu with the following sections:

- Overview
- Activity log
- Tags
- Diagnose and solve problems
- Access Control (IAM)
- Data migration
- Events
- Storage browser (preview)
- Data storage
 - Containers
 - File shares
 - Queues
 - Tables
- Security + networking
 - Networking
 - Azure CDN
 - Access keys
 - Shared access signature
 - Encryption
 - Microsoft Defender for Cloud
- Data management
 - Geo-replication
 - Data protection
 - Object replication

The main content area shows the "strfactsblob" storage account. The "Blob containers" section is expanded, showing "bcontainer1". The "analyticsdata" container is selected, displaying a list of blobs. The authentication method is "Access key (Switch to Azure AD User Account)". The search bar is "Search blobs by prefix (case-sensitive)". The table shows the first 100 items, with columns for "Name" and "Last modified".

Name	Last modified
[.]	
area=antwerp	
area=asheville	
area=athens	
area=austin	
area=bangkok	
area=barcelona	
area=barossa-valley	
area=barwon-south-west-vic	
area=beijing	
area=belize	
area=bergamio	
area=berlin	
area=bologna	
area=bordeaux	
area=boston	
area=bristol	
area=broward-county	

Deployment Architecture



Pipeline Orchestration

Microsoft Azure | Databricks Portal | sxu2012@gmail.com

Workflows > Jobs > short-term-rental-facts-monthly > Run 31163

short-term-rental-facts-monthly run Delete job run

```
graph TD; A["data-acquisition  
Succeeded · 55s  
Python file at dbfs:/FileSto...  
sparkcluster1"] --> B["data-ingestion  
Succeeded · 10m 33s  
Python file at dbfs:/FileSto...  
sparkcluster1"]; B --> C["data-analytics  
Succeeded · 44m 58s  
Python file at dbfs:/FileSto...  
sparkcluster1"];
```

Job run details

Job ID
[490860994334471](#)

Job run ID
31163

Started
2022-06-15 20:21:11 HST

Duration
56m 28s

Status
Succeeded

Clusters

sparkcluster1
Driver: Standard_DS3_v2, Workers: Standard_DS3_v2, 0 workers, 10.4 LTS (includes Apache Spark 3.2.1, Scala 2.12)

View cluster Spark UI Logs Metrics

Pipeline Monitoring

cluster Report at Thu, 16 Jun 2022 07:22:47 +0000

Get Fresh Data

Last

or from to

Go

Clear

Timezone:

Physical View

Grid > cluster > --Choose a Node v

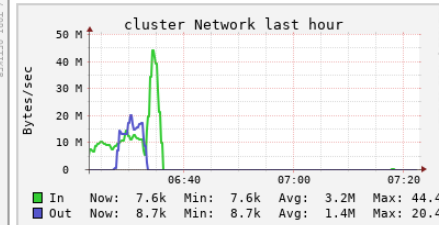
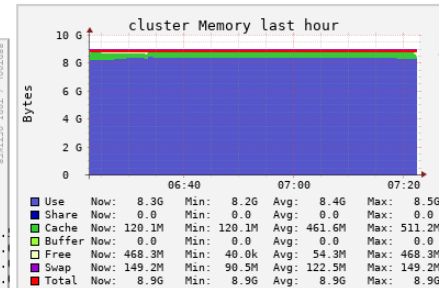
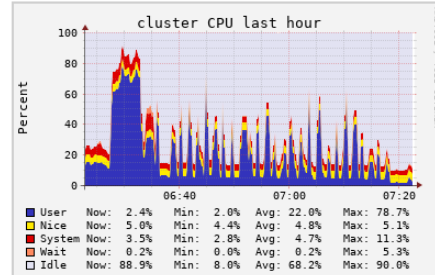
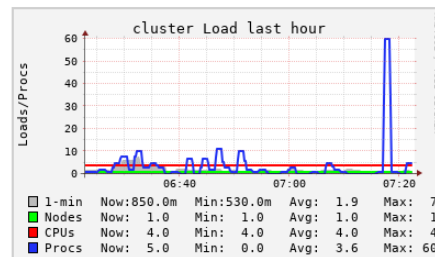
Overview of cluster @ 2022-06-16 07:22

CPUs Total: **4**
Hosts up: **1**
Hosts down: **0**

Current Load Avg (15, 5, 1m):
31%, 25%, 21%

Avg Utilization (last hour):
0%

Server Load Distribution



Stacked Graph - load_one

cluster load_one

cluster load_one last hour sorted by name

Metric

load_one