**Mulai dengan Azure Databricks**

Azure Databricks memungkinkan Anda membangun solusi pemrosesan data dan pembelajaran mesin yang sangat skalabel.

**Tujuan pembelajaran**

Setelah menyelesaikan modul ini, Anda akan dapat:

* Memahami Azure Databricks
* Menyediakan ruang kerja dan kluster Azure Databricks
* Bekerja dengan notebook di Azure Databricks

1. **Pengantar**

Azure Databricks adalah layanan analitik Microsoft, bagian dari platform cloud Microsoft Azure. Ini menawarkan integrasi antara Microsoft Azure dan implementasi Databricks Apache Spark. Azure Databricks terintegrasi secara native dengan keamanan Azure dan layanan data. Dalam modul ini, Anda akan mempelajari cara bekerja dengan fitur utama Azure Databricks.

1. **Tujuan pembelajaran**

Setelah menyelesaikan modul ini, Anda akan mampu:

* Jelaskan konsep utama di Azure Databricks.
* Bekerja dengan ruang kerja dan kluster.
* Bekerja dengan notebook.

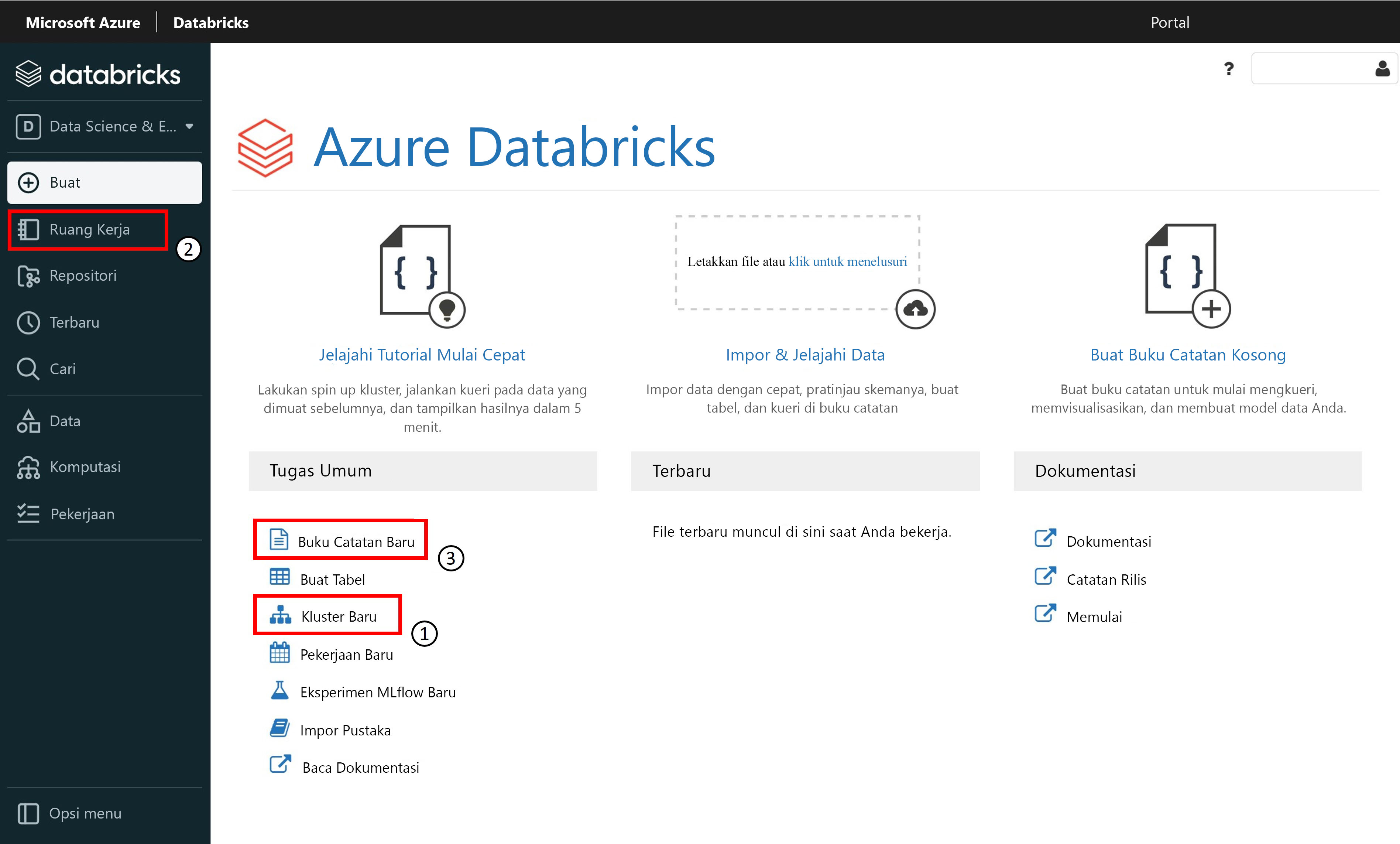
1. **Memahami Azure Databricks**

Azure Databricks berjalan di atas komputer pemrosesan data berpemilik yang disebut Databricks Runtime, versi Apache Spark yang dioptimalkan. Ini memungkinkan kinerja hingga 50x untuk beban kerja Apache Spark.

Apache Spark adalah teknologi inti. Spark adalah mesin analitik sumber terbuka untuk pemrosesan data skala besar. Ini menyediakan antarmuka untuk memprogram seluruh kluster dengan paralelisme data implisit dan toleransi kesalahan.

Singkatnya: Azure Databricks menawarkan layanan analitik berbasis Spark yang cepat, mudah, dan kolaboratif. Ini digunakan untuk mempercepat analitik data besar, kecerdasan buatan, danau data berkinerja tinggi, ilmu data interaktif, pembelajaran mesin, dan kolaborasi.

1. **Konsep utama di Azure Databricks**



Halaman arahan menunjukkan konsep dasar yang akan digunakan di Databricks:

1. **Kluster**: sekumpulan sumber daya komputasi tempat kami menjalankan kode.
2. **Ruang kerja**: mengelompokkan semua elemen Databricks, kluster, buku catatan, data.
3. **Notebook:** dokumen yang berisi kode yang dapat dijalankan, teks deskriptif, dan visualisasi.

**Informasi selengkapnya:** untuk informasi selengkapnya tentang Azure Databricks, lihat dokumentasi.

Link : <https://learn.microsoft.com/id-id/azure/databricks/introduction/>

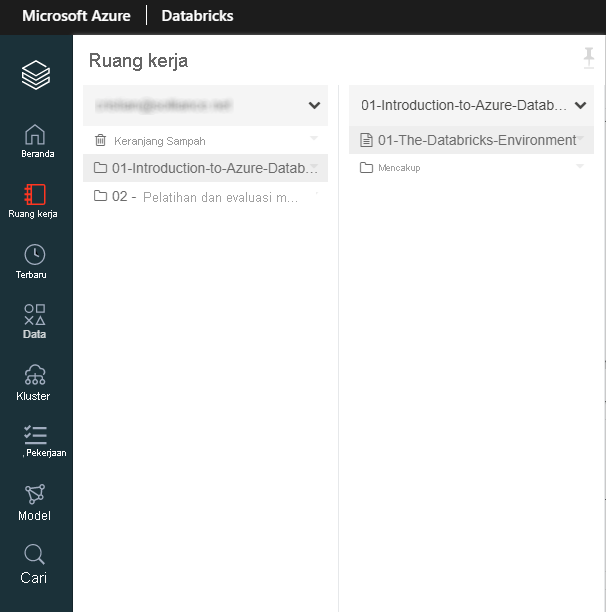
**3. Menyediakan ruang kerja dan kluster Azure Databricks**

Dua konsep utama yang perlu Anda pahami saat bekerja dengan Azure Databricks adalah **ruang kerja** dan **kluster**.

1. **Ruang kerja**

Ruang kerja adalah lingkungan untuk mengakses semua elemen Databricks Anda:

* Ini mengelompokkan objek (seperti buku catatan, pustaka, eksperimen) ke dalam folder,
* Menyediakan akses ke data Anda,
* Menyediakan akses ke sumber daya komputasi yang digunakan (kluster, job).



Setiap pengguna memiliki folder rumah untuk buku catatan dan pustaka mereka. Objek yang disimpan di folder root Ruang kerja adalah: folder, buku catatan, pustaka, dan eksperimen.

Untuk melakukan tindakan pada objek Ruang kerja, kita dapat mengklik kanan objek dan memilih salah satu tindakan yang tersedia.

1. **Kluster**

Kluster adalah sekumpulan sumber daya komputasi tempat Anda menjalankan kode (sebagai buku catatan atau pekerjaan). Kami dapat menjalankan pipeline ETL, atau pembelajaran mesin, ilmu data, beban kerja analitik di kluster.

Kami dapat membuat:

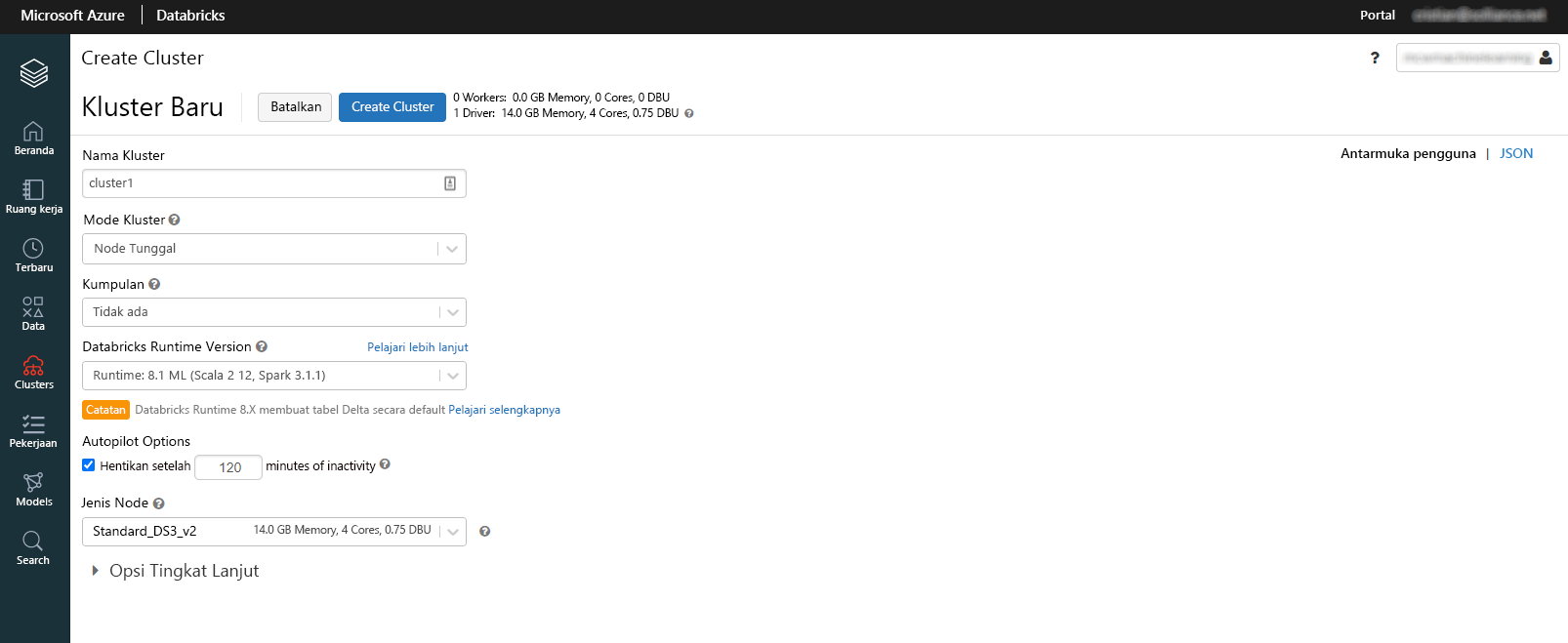
* Sebuah kluster **serba guna**. Beberapa pengguna dapat berbagi kluster tersebut untuk melakukan analisis interaktif kolaboratif.
* **Kluster tugas** untuk menjalankan tugas tertentu. Kluster akan dihentikan ketika pekerjaan selesai (Pekerjaan adalah cara menjalankan notebook atau JAR baik segera atau berdasarkan jadwal).

Sebelum dapat menggunakan kluster, kita harus memilih salah satu **runtime** yang tersedia.

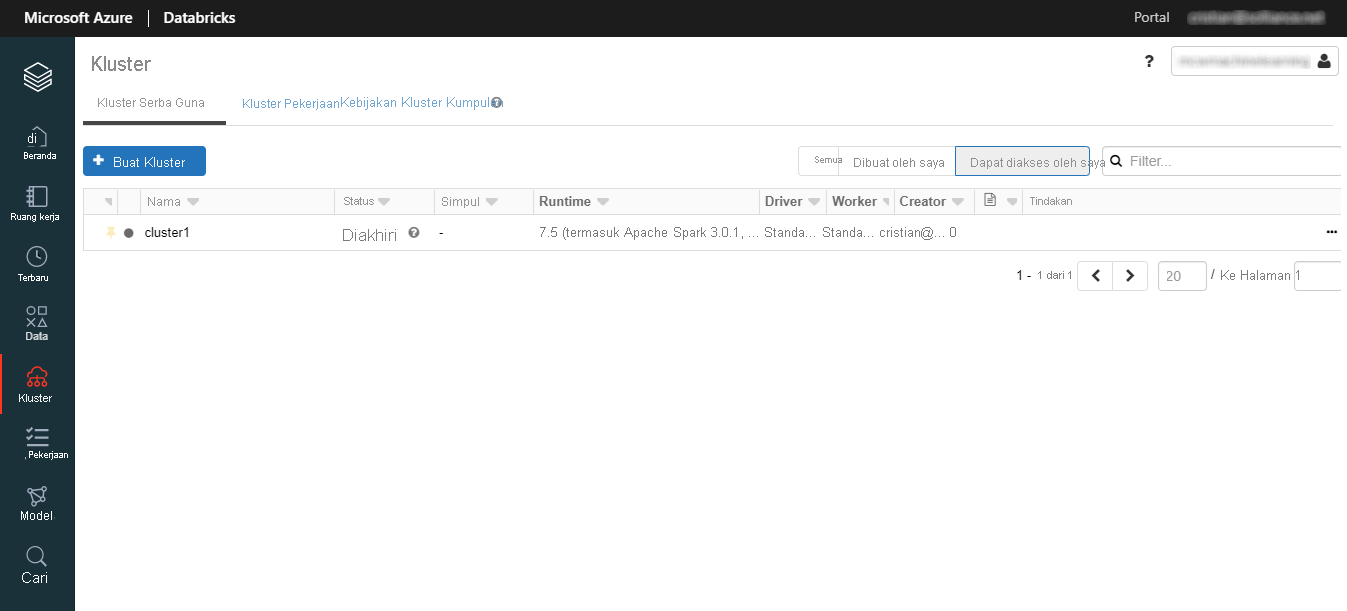
Runtime Databricks adalah kumpulan komponen inti yang berjalan di kluster Azure Databricks. Azure Databricks menawarkan beberapa jenis runtime:

* **Databricks Runtime:** mencakup Apache Spark, komponen, dan pembaruan yang mengoptimalkan kegunaan, kinerja, dan keamanan untuk analitik data besar.
* **Waktu Proses Databricks untuk Pembelajaran Mesin:** varian yang menambahkan beberapa pustaka pembelajaran mesin seperti TensorFlow, Keras, dan PyTorch.
* **Databricks Light:** untuk pekerjaan yang tidak memerlukan kinerja tingkat lanjut, keandalan, atau penskalaan otomatis Databricks Runtime.

Untuk membuat dan mengonfigurasi kluster baru, kita harus memilih tombol **Buat Kluster** dan memilih opsi kita.



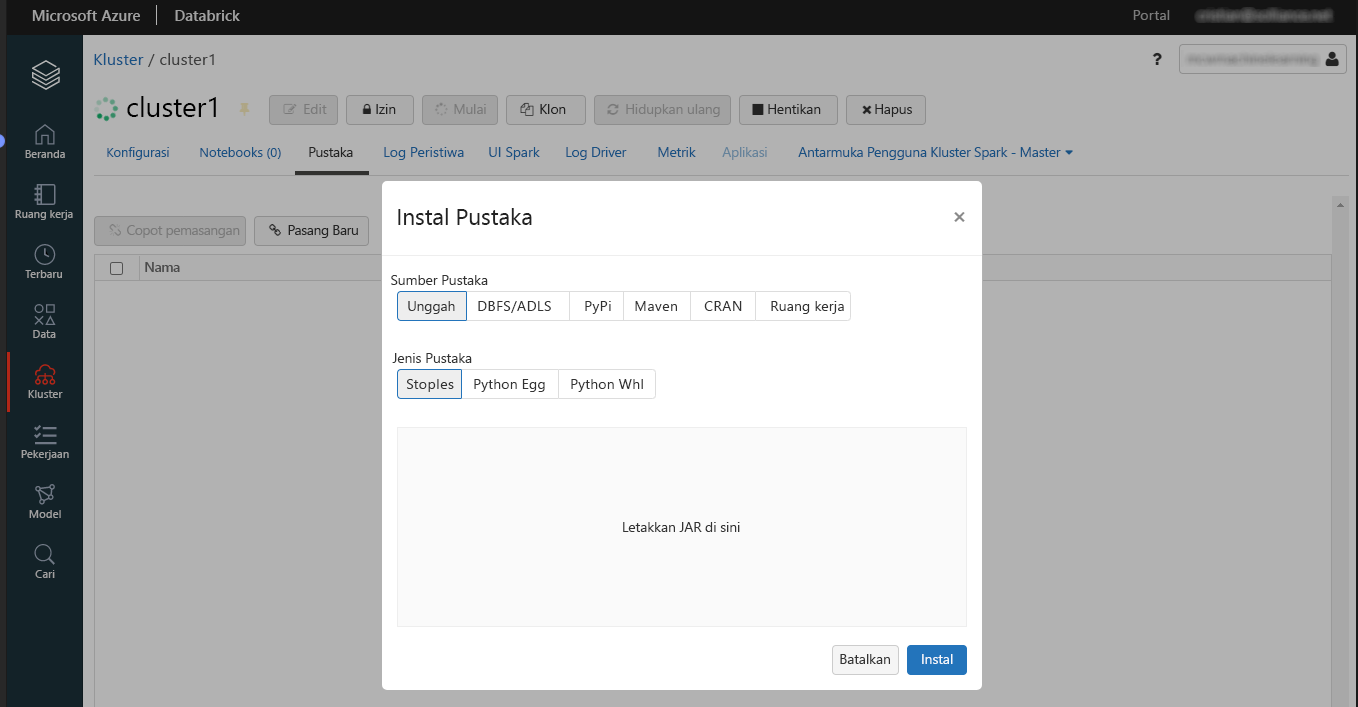
Kami akan melihat kluster baru Anda muncul di daftar kluster.



Untuk meluncurkan kluster, kita harus memilih tombol **Mulai** lalu mengonfirmasi untuk meluncurkannya. Disarankan untuk menunggu sampai kluster dimulai.

Sebuah kluster dapat dikustomisasi dalam banyak cara. Jika Anda ingin membuat kode pihak ketiga tersedia untuk buku catatan Anda, Anda bisa menginstal pustaka. Kluster Anda dapat disediakan untuk menggunakan pustaka Python/Java/Scala/R melalui PyPI atau Maven.

Setelah kluster berjalan, kita dapat memilih **Edit** untuk mengubah propertinya. Jika kami ingin menyediakan kluster Anda dengan pustaka tambahan, kami dapat memilih **Perpustakaan** lalu memilih **Instal Baru.**



Kami dapat memilih pustaka dan nanti akan tersedia untuk digunakan di buku catatan Anda.

**Catatan**

**Informasi selengkapnya:** untuk informasi selengkapnya tentang menginstal pustaka, lihat **Perpustakaan** dalam dokumentasi Azure Databricks.

Link : <https://learn.microsoft.com/id-id/azure/databricks/libraries/>

1. **Bekerja dengan data di ruang kerja**

Database Azure Databricks adalah kumpulan tabel. Tabel Azure Databricks adalah kumpulan data terstruktur.

Kami dapat menyimpan, memfilter, dan melakukan operasi apa pun yang didukung oleh Apache Spark DataFrames pada tabel Azure Databricks. Kita dapat membuat kueri tabel dengan Spark API dan Spark SQL.

Untuk mengakses data kami:

* Kami dapat mengimpor file kami ke DBFS menggunakan UI.
* Kami dapat menginstal dan menggunakan sumber data yang didukung melalui DBFS.

Kami kemudian dapat menggunakan Spark atau API lokal untuk mengakses data.

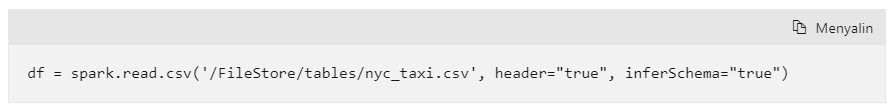
Kami akan dapat menggunakan jalur file DBFS di notebook kami untuk mengakses data kami, terlepas dari sumber datanya.

Dimungkinkan untuk mengimpor data atau kode yang ada di ruang kerja.

Jika kami menggunakan file data kecil di komputer lokal yang ingin kami analisis dengan Azure Databricks, kami dapat mengimpornya ke DBFS menggunakan UI. Ada dua cara untuk mengunggah data ke DBFS dengan UI:

* Unggah file ke FileStore di UI Unggah Data.
* Mengunggah data ke tabel dengan UI Buat tabel, yang juga dapat diakses melalui kotak Impor & Jelajahi Data di halaman arahan.

Kami juga dapat membaca data pada node kluster menggunakan Spark API. Kita dapat membaca data yang diimpor ke DBFS ke dalam Apache Spark DataFrames. Misalnya, jika Anda mengimpor file CSV, Anda dapat membaca data menggunakan kode ini



Kami juga dapat membaca data yang diimpor ke DBFS dalam program yang berjalan pada node driver Spark menggunakan API file lokal. Contohnya:



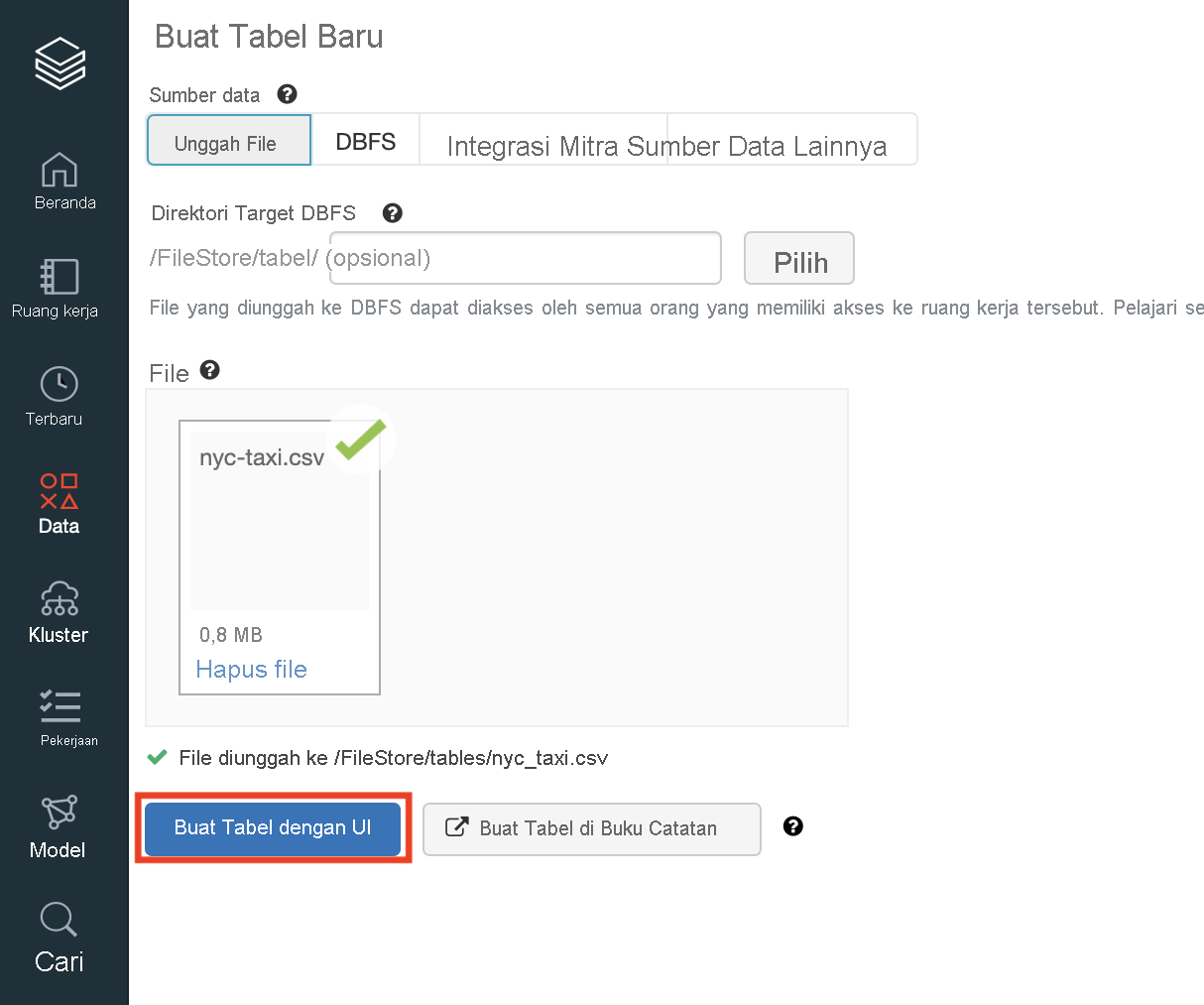
1. **Mengimpor data**

Untuk menambahkan data, kita dapat membuka halaman arahan dan memilih **Impor Jelajahi Data.**

Untuk mendapatkan data dalam tabel, ada beberapa opsi yang tersedia:

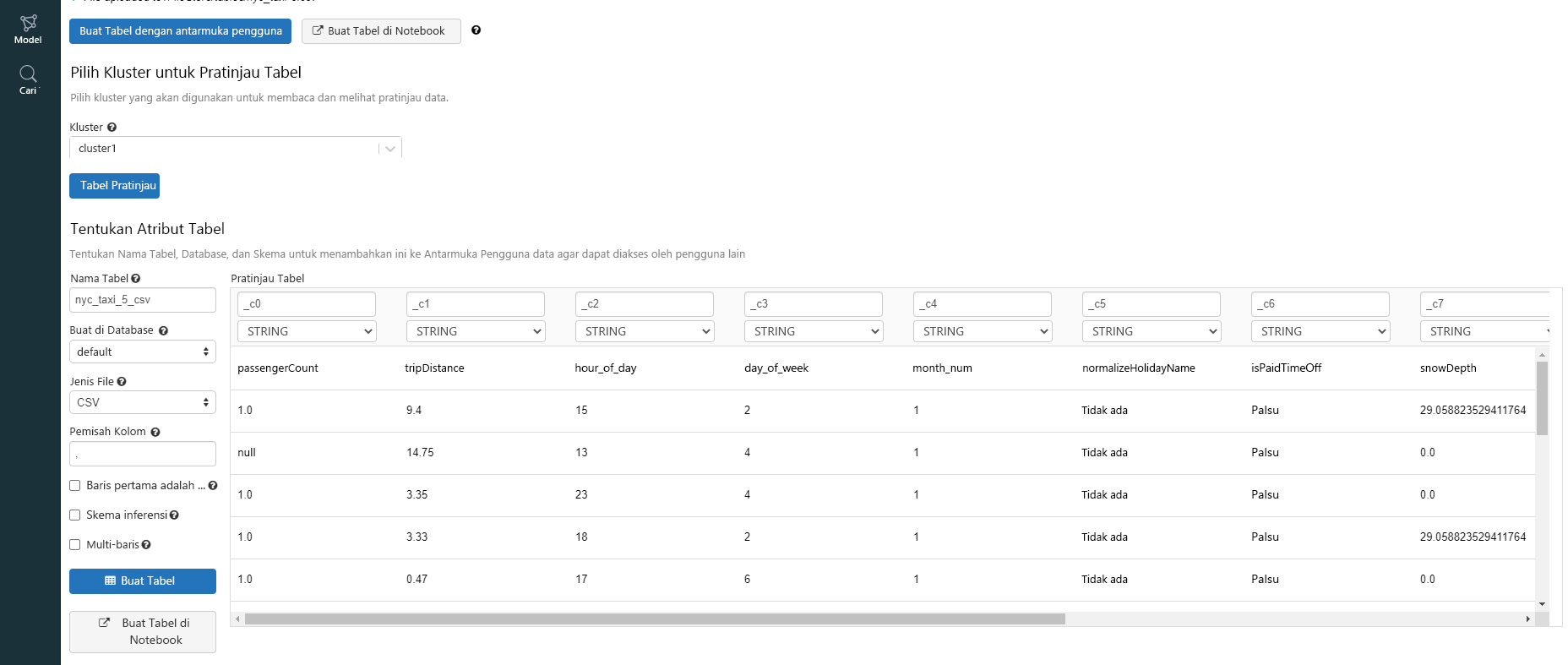
* Unggah file lokal dan impor datanya.
* Gunakan data yang sudah ada di bagian DBFS.
* Instal sumber data eksternal, seperti Azure Storage, Azure Data Lake, dan lainnya.

Untuk membuat tabel berdasarkan file lokal, kami dapat memilih **Unggah File** untuk mengunggah data dari komputer lokal Anda.

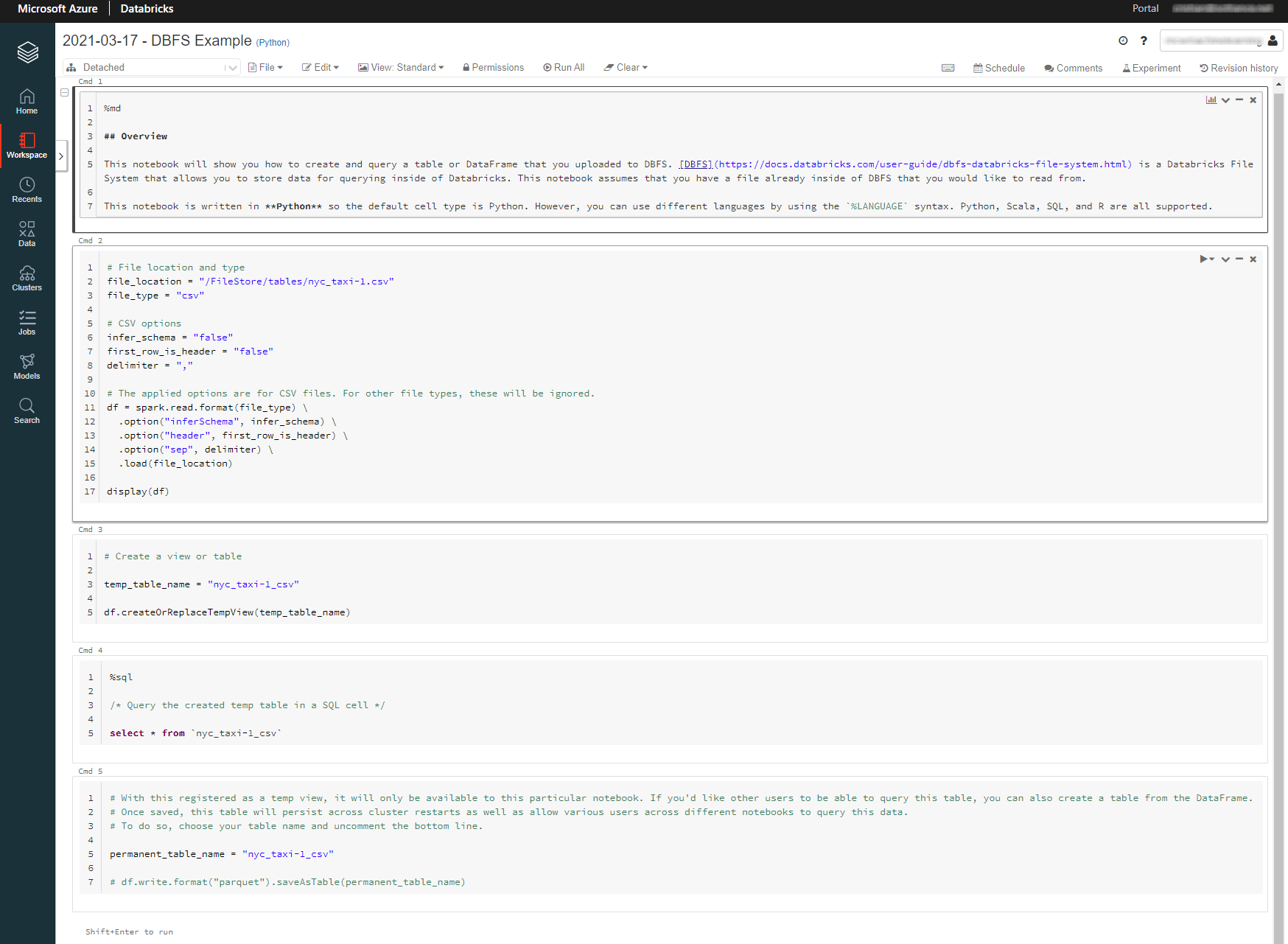


Setelah data diunggah, itu akan tersedia sebagai tabel atau sebagai titik mount di bagian sistem file DBFS **(/FileStore).**

Databricks dapat membuat tabel secara otomatis jika kita memilih **Buat Tabel dengan UI.**



Sebagai alternatif, kita dapat memiliki kontrol penuh atas struktur tabel baru dengan memilih **Buat Tabel di Notebook**. Azure Databricks akan menghasilkan kode Spark yang memuat data Anda (dan kami dapat menyesuaikannya melalui API Spark).



1. **Menggunakan data terinstal DBFS**

Databricks File System (DBFS) adalah sistem file terdistribusi yang diinstal ke ruang kerja Databricks dan tersedia di kluster Databricks. DBFS adalah abstraksi di atas penyimpanan objek yang dapat diskalakan dan menawarkan manfaat berikut:

* Memungkinkan Anda memasang objek penyimpanan sehingga Anda dapat mengakses data dengan lancar tanpa memerlukan info masuk.
* Memungkinkan Anda berinteraksi dengan penyimpanan objek menggunakan direktori dan semantik file, bukan URL penyimpanan.
* Menyimpan file ke penyimpanan objek, sehingga Anda tidak akan kehilangan data setelah menghentikan kluster.

Lokasi penyimpanan default di DBFS dikenal sebagai root DBFS.

Kita dapat menggunakan DBFS untuk mengakses:

* File lokal (sebelumnya diimpor). Misalnya, tabel yang Anda impor di atas tersedia di bagian **/FileStore**
* File jarak jauh, objek yang disimpan di penyimpanan terpisah seolah-olah berada di sistem file lokal

Misalnya, untuk menginstal akun penyimpanan Azure jarak jauh sebagai folder DBFS, kita dapat menggunakan modul **dbutils:**

data\_storage\_account\_name = '<data\_storage\_account\_name>'

data\_storage\_account\_key = '<data\_storage\_account\_key>'

data\_mount\_point = '/mnt/data'

data\_file\_path = '/bronze/wwi-factsale.csv'

dbutils.fs.mount(

source = f"wasbs://dev@{data\_storage\_account\_name}.blob.core.windows.net",

mount\_point = data\_mount\_point,

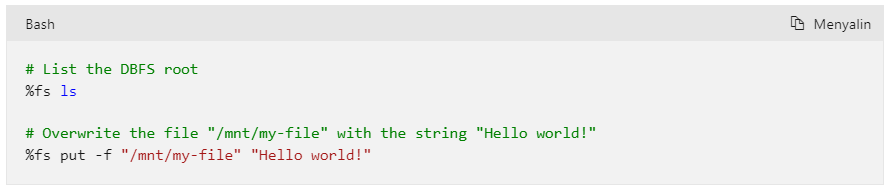
extra\_configs = {f"fs.azure.account.key.{data\_storage\_account\_name}.blob.core.windows.net": data\_storage\_account\_key})

display(dbutils.fs.ls("/mnt/data"))

#this path is available as dbfs:/mnt/data for spark APIs, e.g. spark.read

#this path is available as file:/dbfs/mnt/data for regular APIs, e.g. os.listdir

Notebook mendukung steno - **%fs** perintah ajaib - untuk mengakses modul sistem file dbutils. Sebagian besar perintah **dbutils.fs** tersedia menggunakan **%fs** perintah ajaib:



**Catatan**

**Informasi selengkapnya:** untuk informasi selengkapnya tentang DBFS, lihat **Sistem File Databricks** dalam dokumentasi Azure Databricks.

Link : <https://learn.microsoft.com/id-id/azure/databricks/dbfs/>

**4. Bekerja dengan notebook di Azure Databricks**

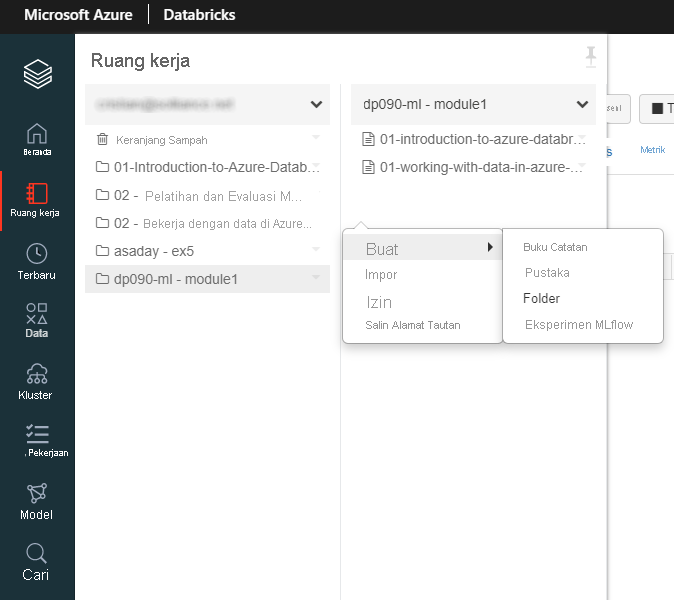
Notebook adalah antarmuka berbasis web ke dokumen yang berisi:

* Kode yang dapat dijalankan
* Teks deskriptif
* Visualisasi

Notebook adalah kumpulan sel yang dapat dijalankan (perintah). Saat Anda menggunakan buku catatan, Anda terutama mengembangkan dan menjalankan sel.

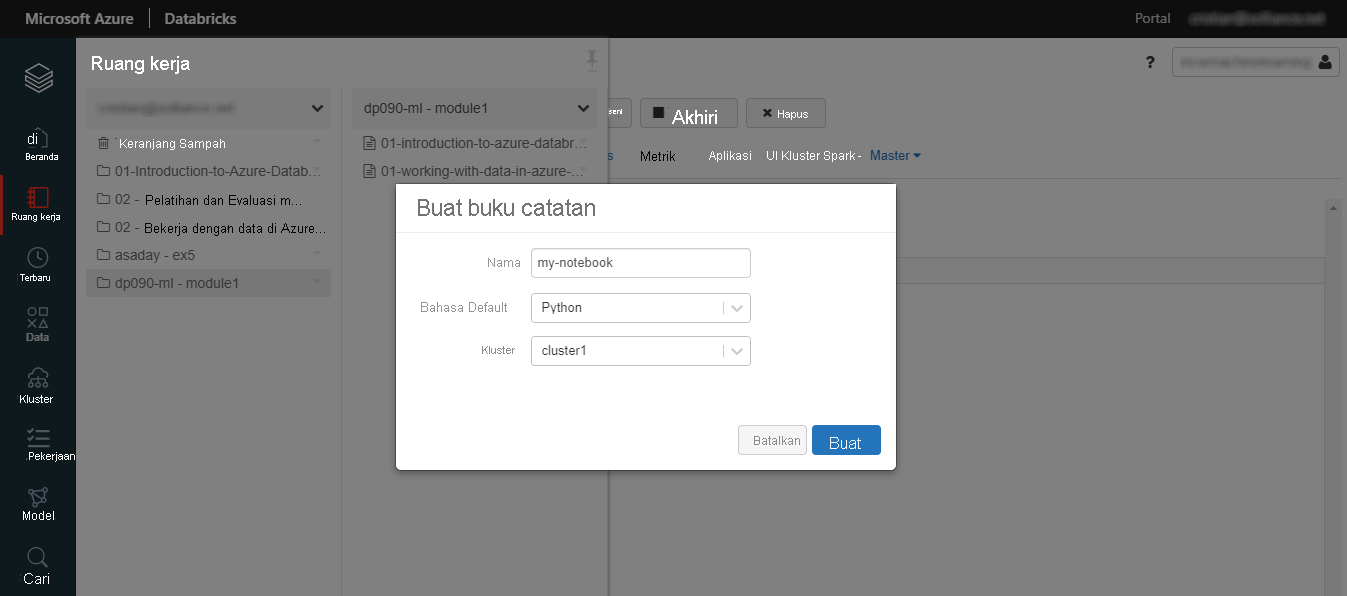
Sel yang dapat dijalankan beroperasi pada file dan tabel. Sel dapat dijalankan secara berurutan, mengacu pada output dari sel yang dijalankan sebelumnya.

Untuk membuat buku catatan, kita dapat memilih **Ruang Kerja**, menelusuri folder yang diinginkan, klik kanan, dan pilih **Buat**, lalu pilih **Buku Catatan.**

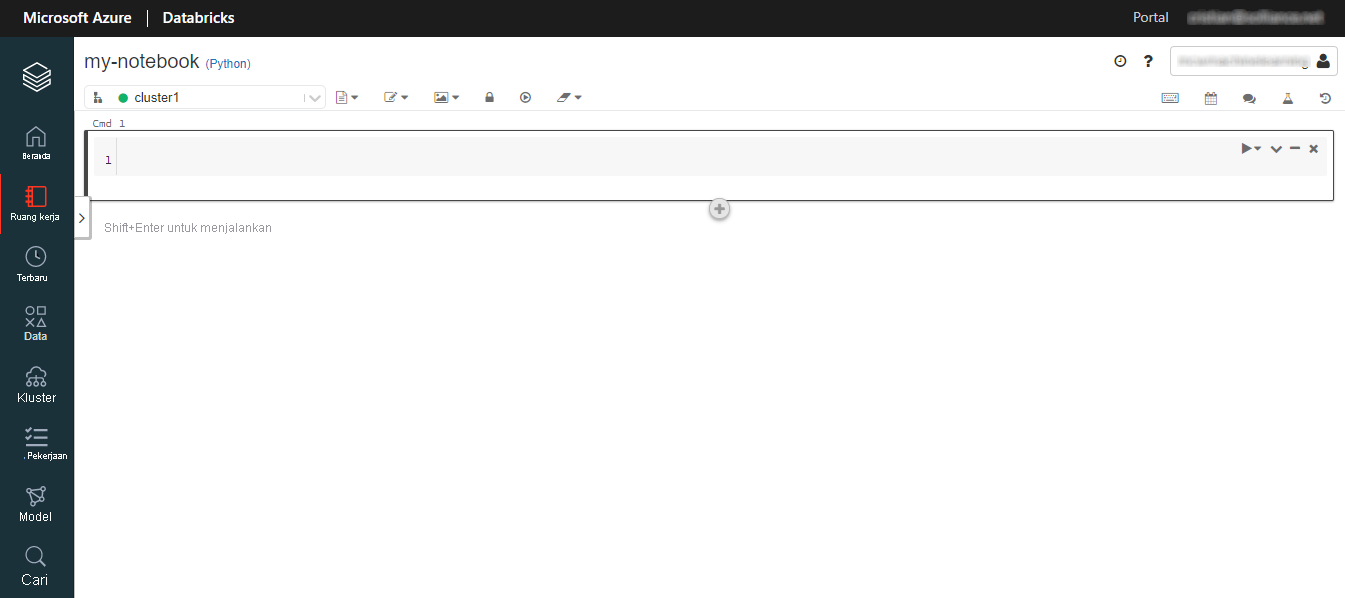


Nama harus diberikan ke buku catatan baru, dan bahasa default untuk digunakan di dalam sel kode. Pilih kluster untuk menjalankan kode dalam sel.

Untuk sel yang dapat dijalankan, bahasa pemrograman berikut didukung: Python, Scala, R, dan SQL. Anda dapat memilih bahasa default untuk sel dalam buku catatan. Anda juga dapat mengganti bahasa itu nanti.



Editor buku catatan terbuka dengan sel kosong pertama



Dengan mengarahkan kursor ke tombol **Plus** di bagian sel saat ini atau dengan memilih opsi menu kanan atas, kita dapat mengubah konten buku catatan. Kami dapat menambahkan sel baru, memotong, menyalin, mengekspor konten sel, atau menjalankan sel tertentu.

Kami dapat mengganti bahasa default dengan menetapkan perintah ajaib bahasa **%<language>** di awal sel.

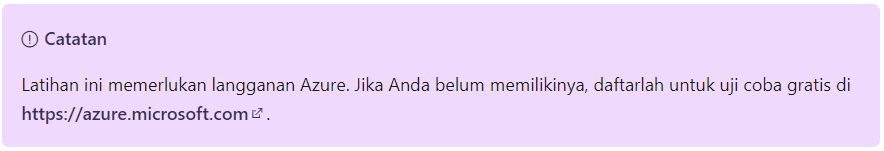
Perintah ajaib yang didukung adalah:

* **%python**
* **%r**
* **%scala**
* **%sql**

Notebook juga mendukung beberapa perintah ajaib tambahan:

* **%sh:** Memungkinkan Anda menjalankan kode shell di buku catatan Anda
* **%fs**: Memungkinkan Anda menggunakan perintah sistem file dbutils
* **%md**: Memungkinkan Anda menyertakan berbagai jenis dokumentasi, termasuk teks, gambar, dan rumus serta persamaan matematika.

**5. Latihan - Mulai dengan Azure Databricks**



Sekarang adalah kesempatan Anda untuk memulai Azure Databricks sendiri dengan mengonfigurasi kluster, membuat ruang kerja, dan buku catatan.

Dalam latihan ini, Anda akan:

* Buat Kluster Azure Databricks.
* Sediakan Ruang Kerja Azure Databricks.
* Bekerja dengan Notebook.
* Gunakan DBFS.

1. **Instruksi**

Ikuti petunjuk berikut untuk menyelesaikan latihan:

* Buka petunjuk latihan di <https://aka.ms/mslearn-dp090.> \ <https://microsoftlearning.github.io/dp-090-databricks-ml/>
* Selesaikan latihan Memulai Azure Databricks.

1. **Ringkasan**

Dalam modul ini, Anda telah mempelajari cara memulai Azure Databricks.

Sekarang setelah Anda menyelesaikan modul ini, Anda dapat:

* Jelaskan konsep utama di Azure Databricks.
* Bekerja dengan ruang kerja dan kluster.
* Bekerja dengan notebook.