Buenas prácticas con Databricks

Cómo usar de la mejor manera Databricks para nuestros procesos de Big Data.

Quarantine Meetup

Mayo 2020

The better the question. The better the answer.
The better the world works.

0

- Néstor Campos
- Arquitecto Senior en EY
- Experiencia y certificado en muchas tecnologías, incluyendo Azure y una especialización en Databricks.
- https://www.linkedin.com/in/nescampos/

1 Parte teórica

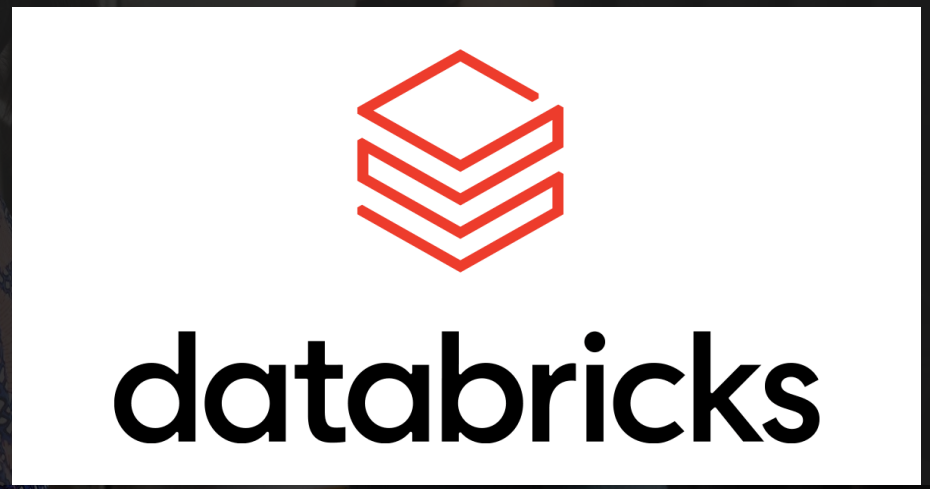
Qué es Databricks

¿Qué es?

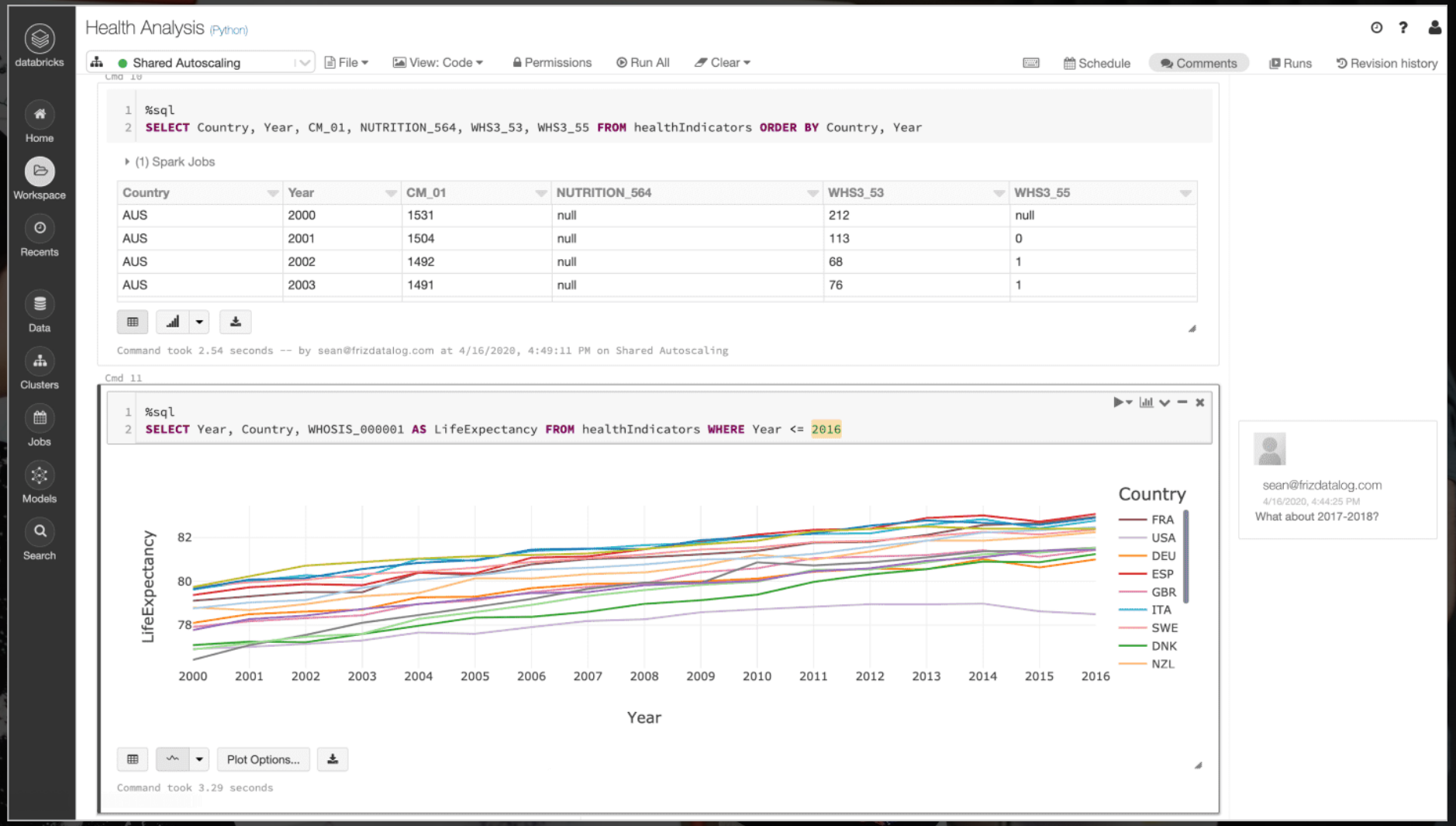
- Un servicio para procesar datos tanto en tiempo real como batch, aprovechando la potencialidad de Spark y múltiples componentes desarrollados específicamente para esta tecnología.

- En términos simples, Spark como servicio.

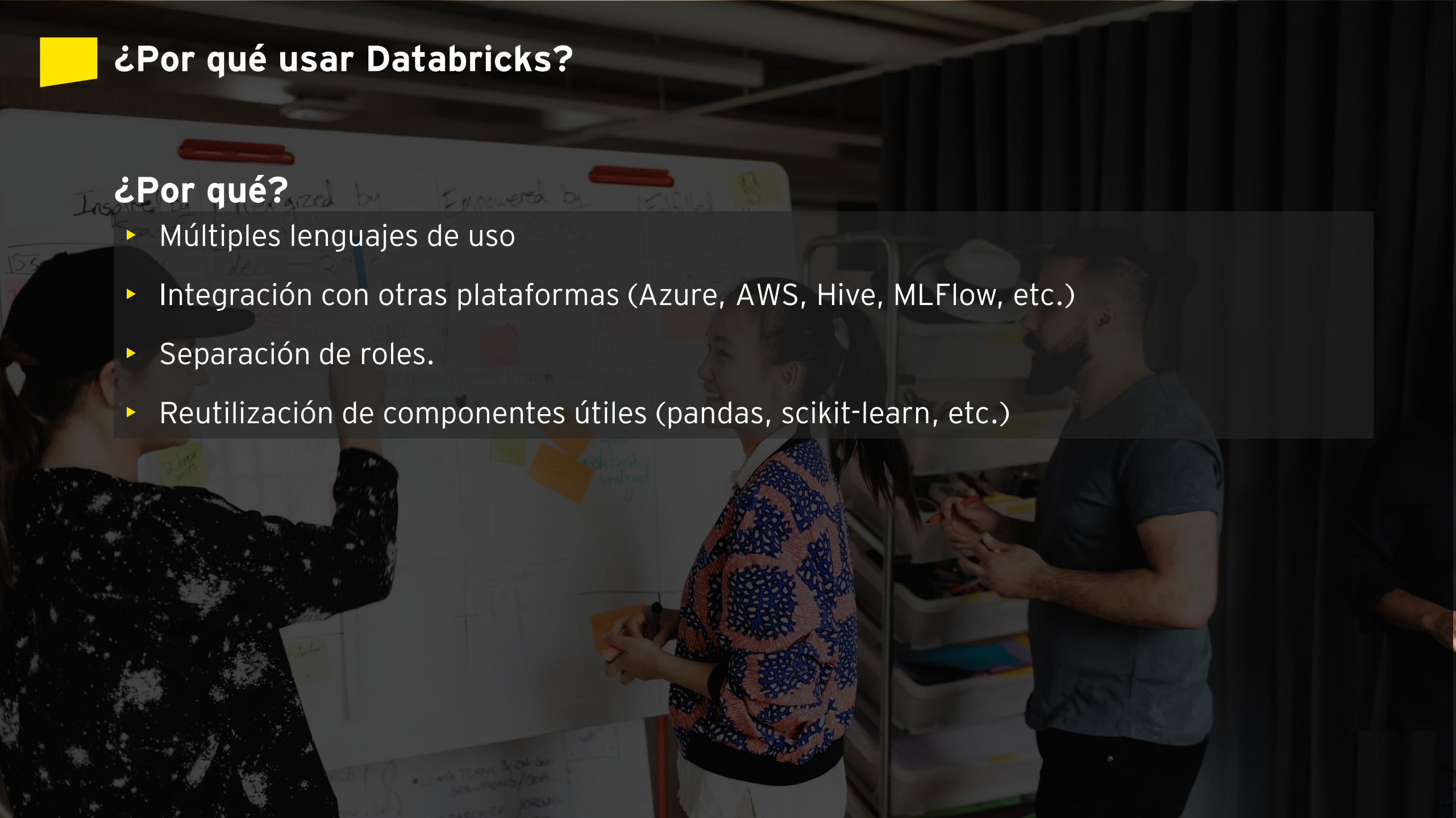
- Link oficial: https://databricks.com/



Databricks


databricks
Home
Workspace
Recents
Data
Clusters
Jobs
Models
Search
Health Analysis (Python)
Shared Autoscaling
File
View: Code
Permissions
Run All
Clear
Schedule
Comments
Runs
Revision history
1 %sql
2 SELECT Country, Year, CM_01, NUTRITION_564, WHS3_53, WHS3_55 FROM healthIndicators ORDER BY Country, Year
(1) Spark Jobs
Country Year CM_01 NUTRITION_564 WHS3_53 WHS3_55
AUS 2000 1531 null 212 null
AUS 2001 1504 null 113 0
AUS 2002 1492 null 68 1
AUS 2003 1491 null 76 1
Command took 2.54 seconds -- by sean@frizdatalog.com at 4/16/2020, 4:49:11 PM on Shared Autoscaling
Cmd 11
1 %sql
2 SELECT Year, Country, WHOSIS_000001 AS LifeExpectancy FROM healthIndicators WHERE Year <= 2016
LifeExpectancy
82
80
78
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016
Year
Country
FRA
USA
DEU
ESP
GBR
ITA
SWE
DNK
NZL
Plot Options...
Command took 3.29 seconds
sean@frizdatalog.com
4/16/2020, 4:44:25 PM
What about 2017-2018?

¿Por qué usar Databricks?



¿Por qué?

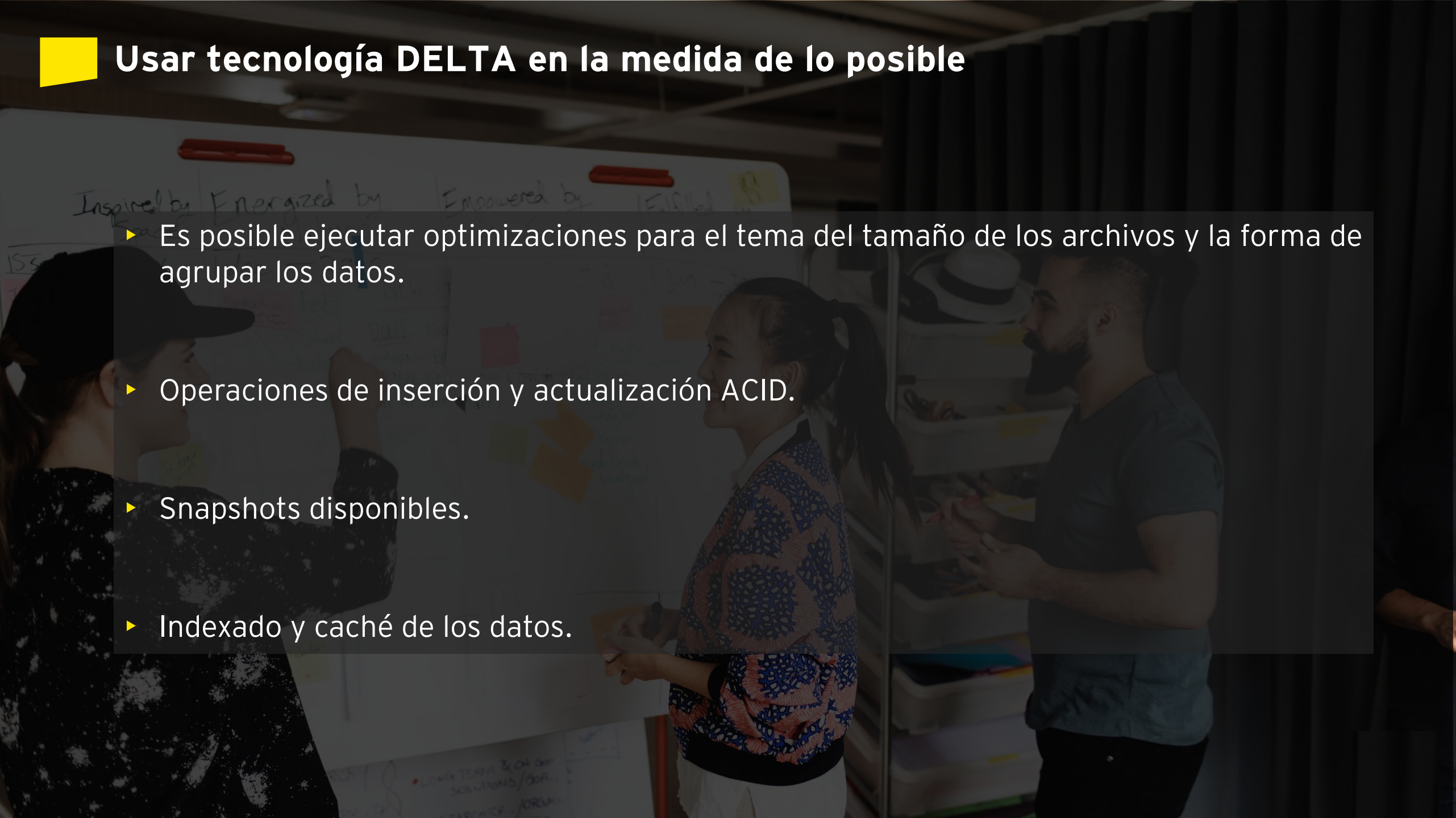
- Múltiples lenguajes de uso
- Integración con otras plataformas (Azure, AWS, Hive, MLFlow, etc.)
- Separación de roles.
- Reutilización de componentes útiles (pandas, scikit-learn, etc.)

2 Buenas prácticas

Separar cada tarea en sus respectivos componentes

- Azure Active Directory: Autenticación y privilegios
- Azure Key Vault: Secretos
- Azure Datalake Storage Gen 2

Usar tecnología DELTA en la medida de lo posible



- Es posible ejecutar optimizaciones para el tema del tamaño de los archivos y la forma de agrupar los datos.
- Operaciones de inserción y actualización ACID.
- Snapshots disponibles.
- Indexado y caché de los datos.

2 Y más que veremos ahora!!!

3 Preguntas


GRACIAS!!!!