Cell Phone Data for Social Impact

MIT CDR Data Training

JAN 2020

MITGOV/LAB

CIVIC DESIGN DATA LAB

DIRECTORATE OF SCIENCE TECHNOLOGY & INNOVATION

africell

Agenda

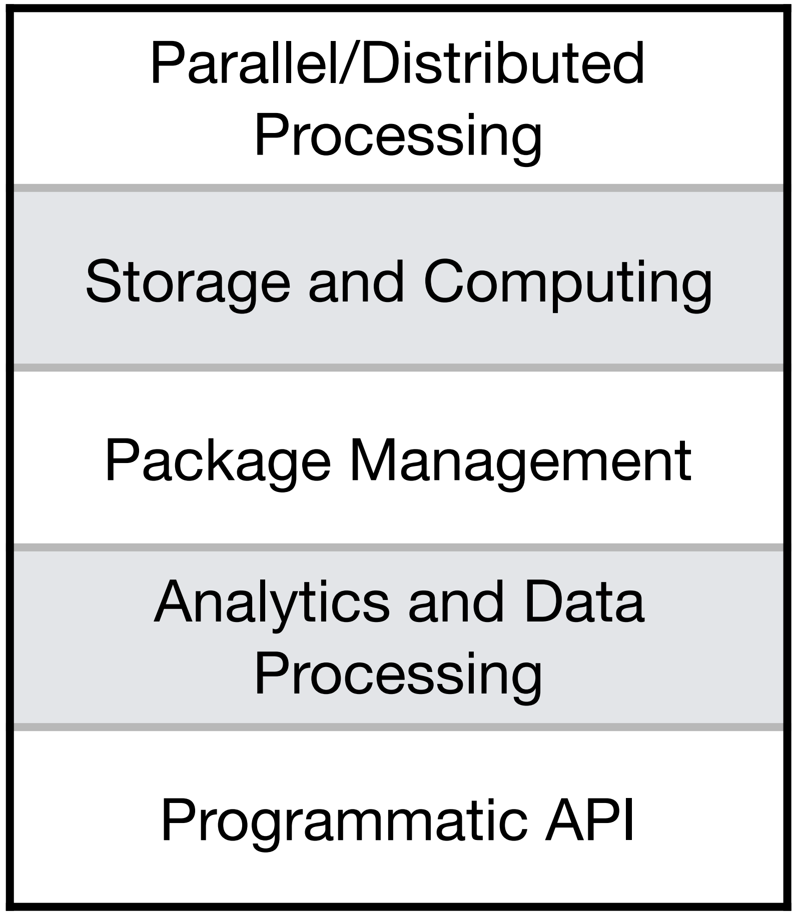
11:00 - 11:30 AM	Preparation
11:30 AM - 12:00 PM	Introduction to Spark
12:00 - 12:15 PM	Break
12:15 - 12:30 PM	Data Preparation
12:30 - 1:15 PM	Data Processing and Exploration
1:15 - 1:30 PM	Closing & Conclusion

MITGOV/LAB CIVIC DATA DESIGN LAB DIRECTORATE OF SCIENCE TECHNOLOGY & INNOVATION africell

Goals & Objectives

1. Understand some basic tools for large-scale, big data processing
2. Wrangle and ready CDR data for processing
3. Learn the basics of Apache Spark
4. Explore the use of Apache Zeppelin for integrated analytics
5. Work collaboratively in a secure analytics environment

Tools & Frameworks


Parallel/Distributed Processing

Storage and Computing

Package Management

Analytics and Data Processing

Programmatic API

APACHE Spark™

aws

ANACONDA®

Apache Zeppelin

MITGOV/LAB

CIVIC DESIGN DATA LAB

DIRECTORATE OF SCIENCE TECHNOLOGY & INNOVATION

africell

X

Setting Up

MITGOV/LAB CIVIC DATA DESIGN LAB DIRECTORATE OF SCIENCE TECHNOLOGY & INNOVATION africell

X Setting Up

1. Install FoxyProxy
2. Download the FoxyProxy settings file
3. Import the settings file to Foxy Proxy
4. Select the proxy setting:
    1. `Use proxy emr-socks-proxy for all URLs`
5. Open your Terminal/Shell and enter the following:
    1.
    ```
    ssh -i ~/.ssh/<private-key> -ND 8157
    <username>@34.223.103.224
    ```
6. Return to your browser

MITGOV/LAB

CIVIC DESIGN DATA LAB

DIRECTORATE OF SCIENCE TECHNOLOGY & INNOVATION

africell

X Setting Up: Why?

Our instance is not accessible via the public internet. This is good. In order to access our analytics environment, you need secure (authenticated and authorized).

We have two basic layers of security: SSH and AWS IAMs

We have control over who has access to CDRs and know what they are doing with it!

I

Introduction to Apache Spark

MITGOV/LAB

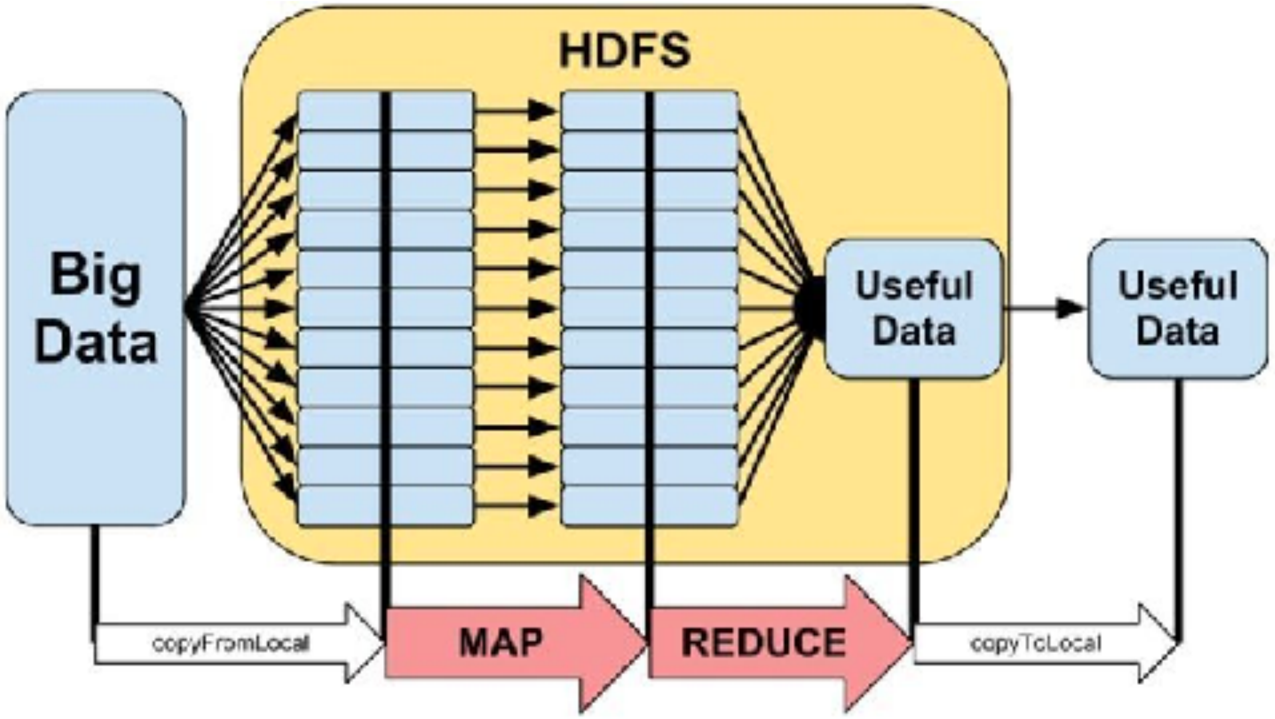
CIVIC DESIGN DATA LAB

DIRECTORATE OF SCIENCE TECHNOLOGY & INNOVATION

africell

I

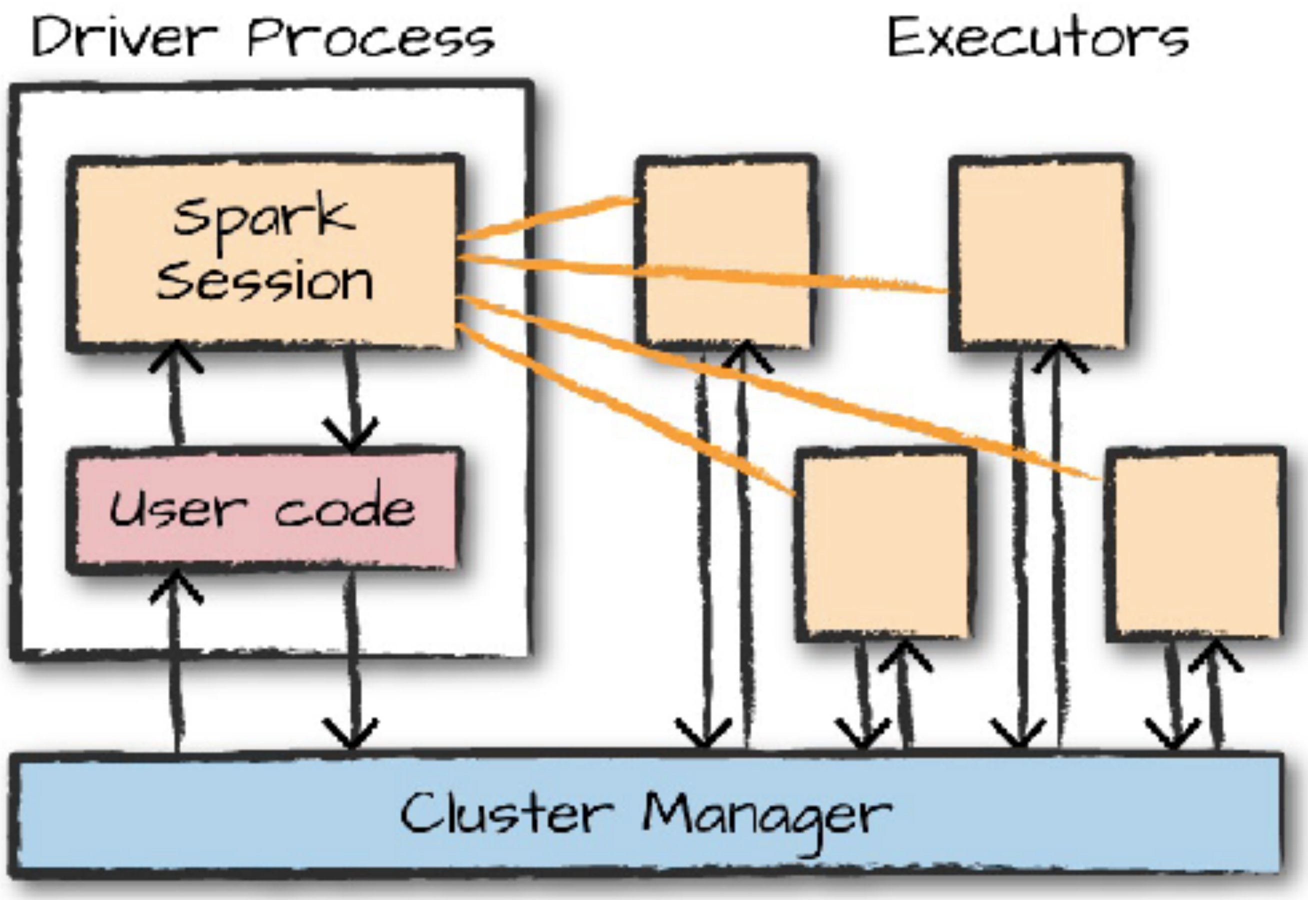
What is Spark?



a better Hadoop?

MITGOV/LAB CIVIC DESIGN DATA LAB DIRECTORATE OF SCIENCE TECHNOLOGY & INNOVATION africell

I What is Spark?



Source: Chamber, Bill et al. 2017. Spark: The Definitive Guide, pg 21

- Spark processes in-memory and can run onto various filesystems (HDFS, S3, RDBMs, etc).
- Very powerful APIs and abstraction.
- Spark is lazy.

I

What is Spark?

Follow on Zeppelin

I What is Spark: Dataframe

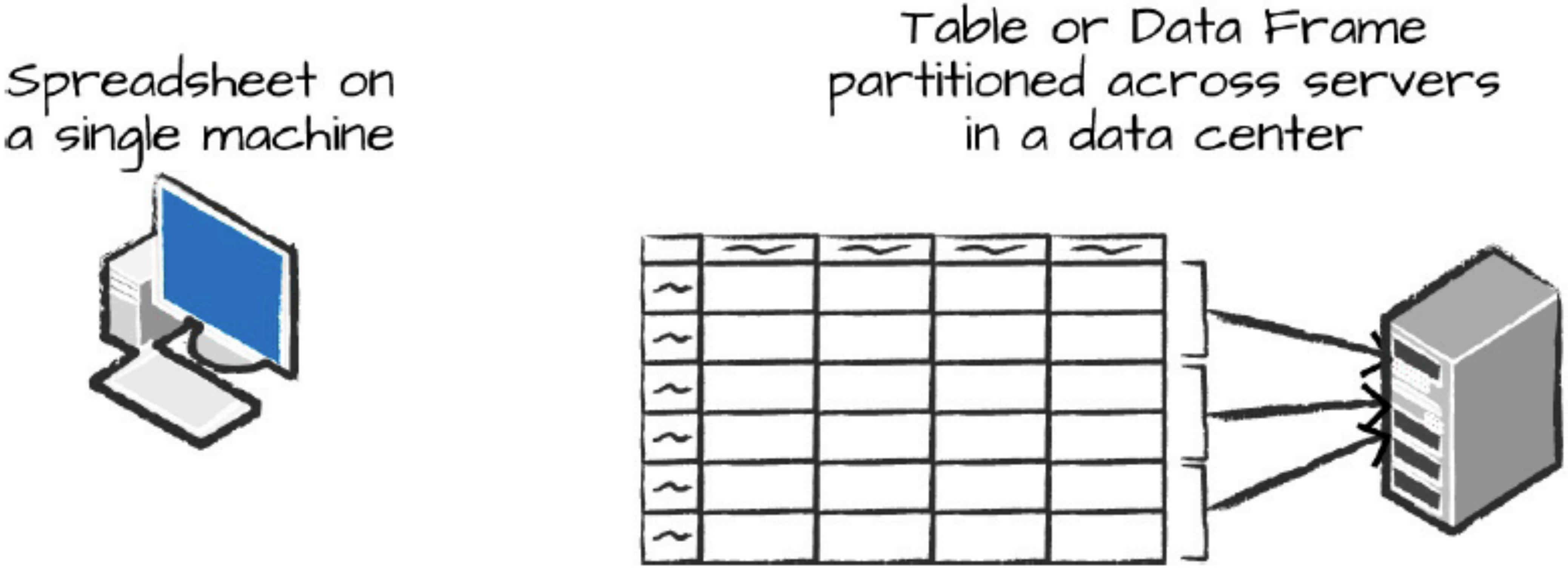

Spreadsheet on a single machine
Table or Data Frame partitioned across servers in a data center

Figure 2-3. Distributed versus single-machine analysis

Source: Chamber, Bill et al. 2017. Spark: The Definitive Guide, pg 25

I

What is Spark: Transformations

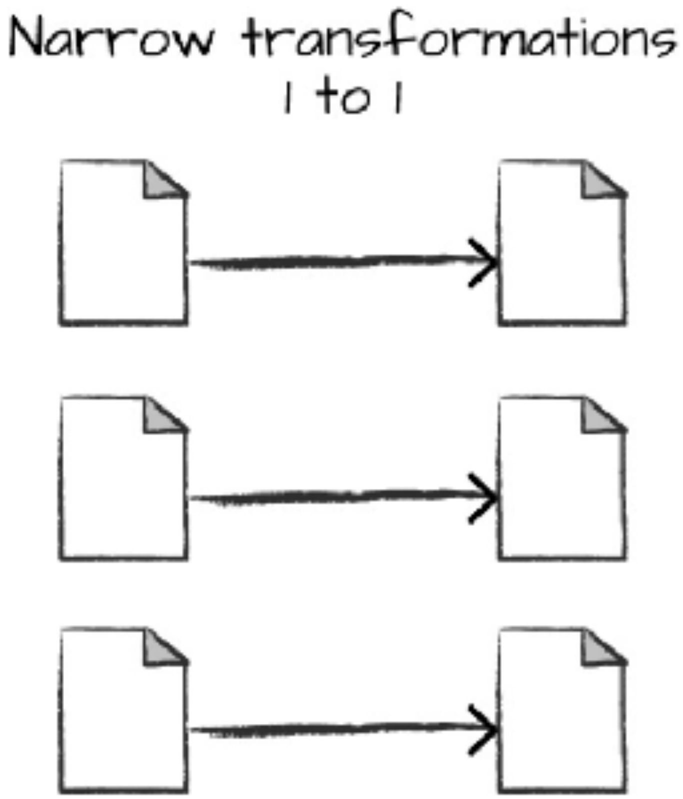


Figure 2-4. A narrow dependency

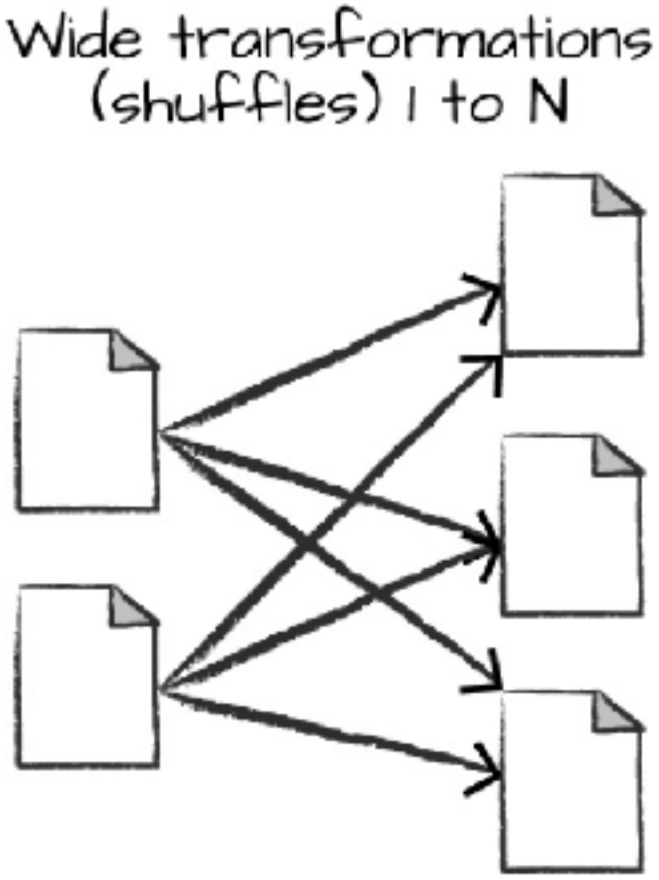


Figure 2-5. A wide dependency

Source: Chamber, Bill et al. 2017. Spark: The Definitive Guide, pg 27-28

I What is Spark: Transformations

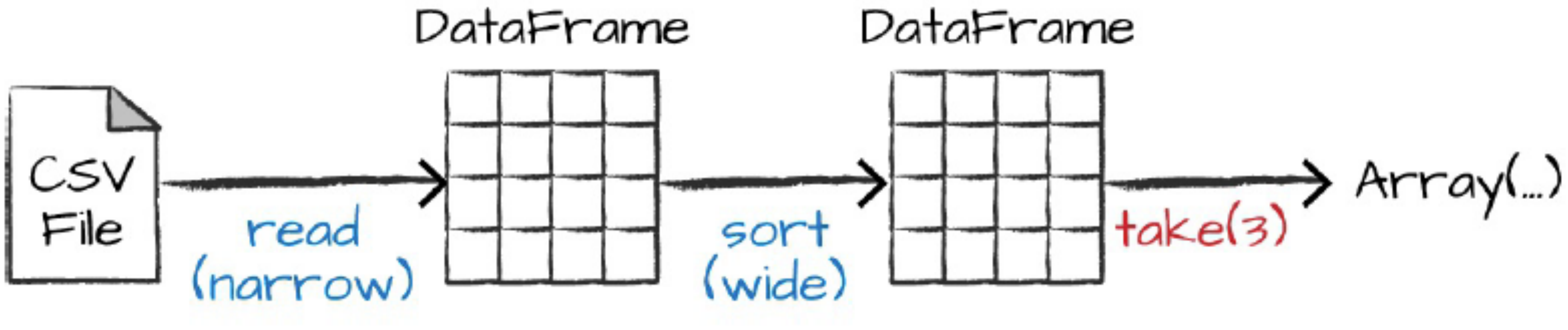


Figure 2-8. Reading, sorting, and collecting a DataFrame

Source: Chamber, Bill et al. 2017. Spark: The Definitive Guide, pg 31

I

What is Spark: Transformations

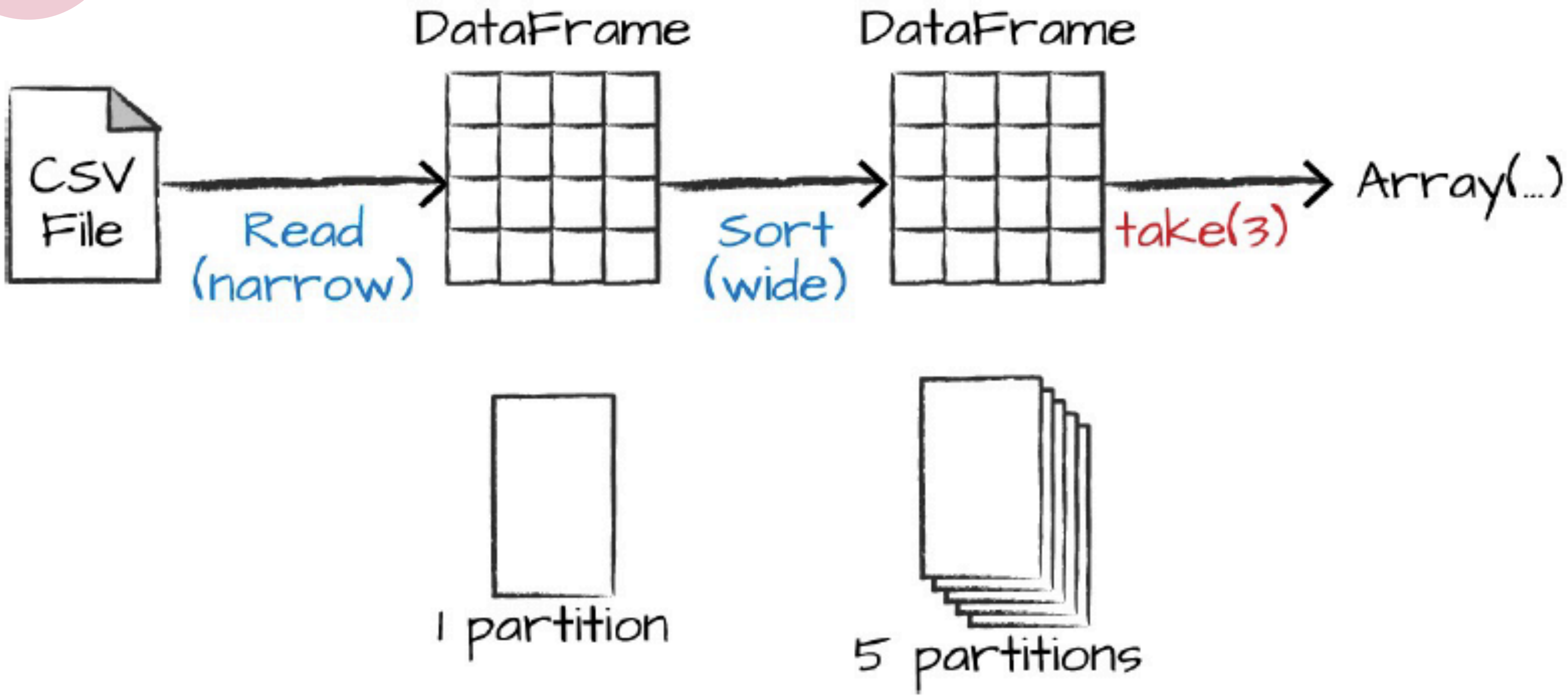


Figure 2-9. The process of logical and physical DataFrame manipulation

Source: Chamber, Bill et al. 2017. Spark: The Definitive Guide, pg 32

MITGOV/LAB CIVIC DESIGN DATA LAB DIRECTORATE OF SCIENCE TECHNOLOGY & INNOVATION africell

I

What is Spark: Transformations

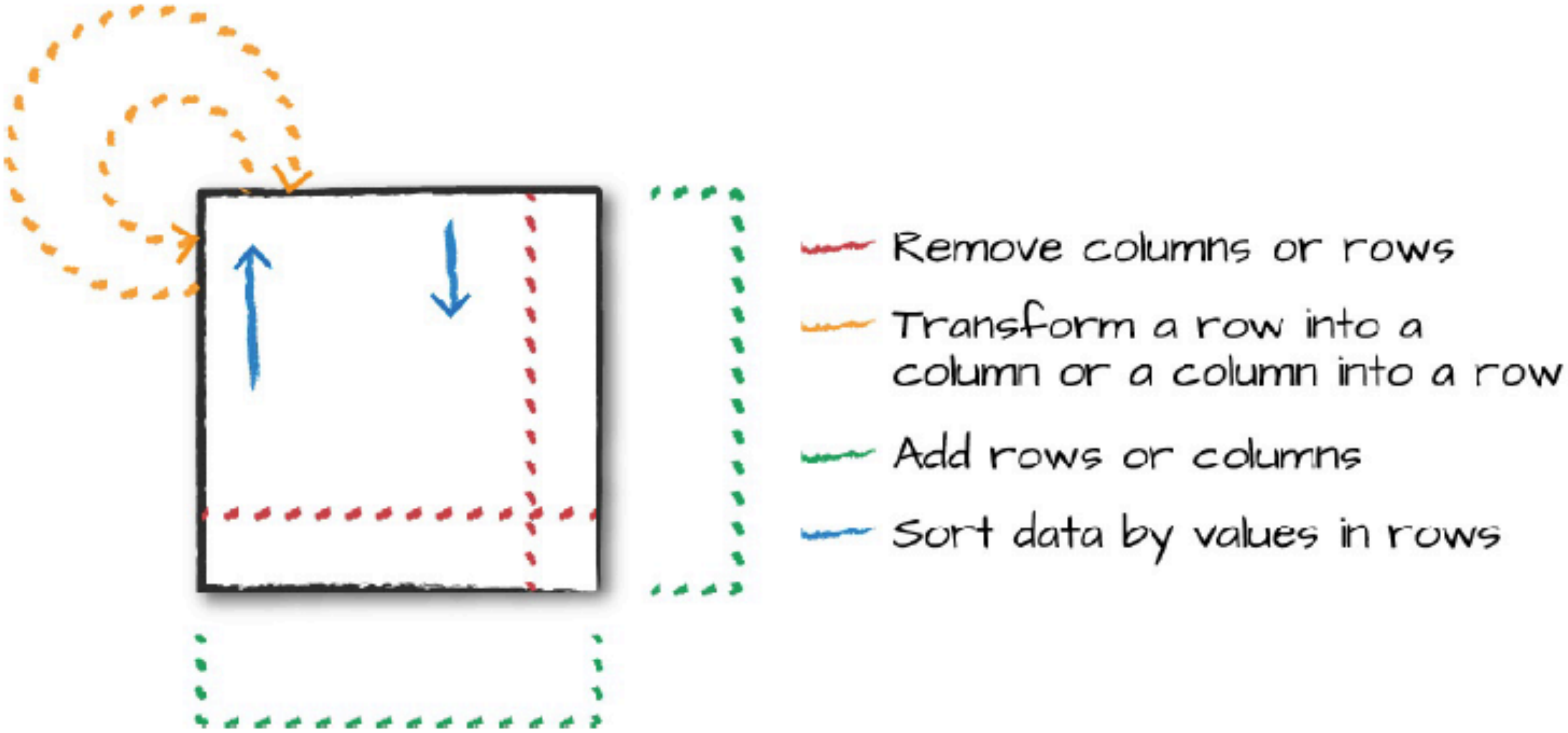

Remove columns or rows
Transform a row into a column or a column into a row
Add rows or columns
Sort data by values in rows

Figure 5-2. Different kinds of transformations

Source: Chamber, Bill et al. 2017. Spark: The Definitive Guide, pg 32

MITGOV/LAB CIVIC DESIGN DATA LAB DIRECTORATE OF SCIENCE TECHNOLOGY & INNOVATION africell

I

What is Spark: Windows

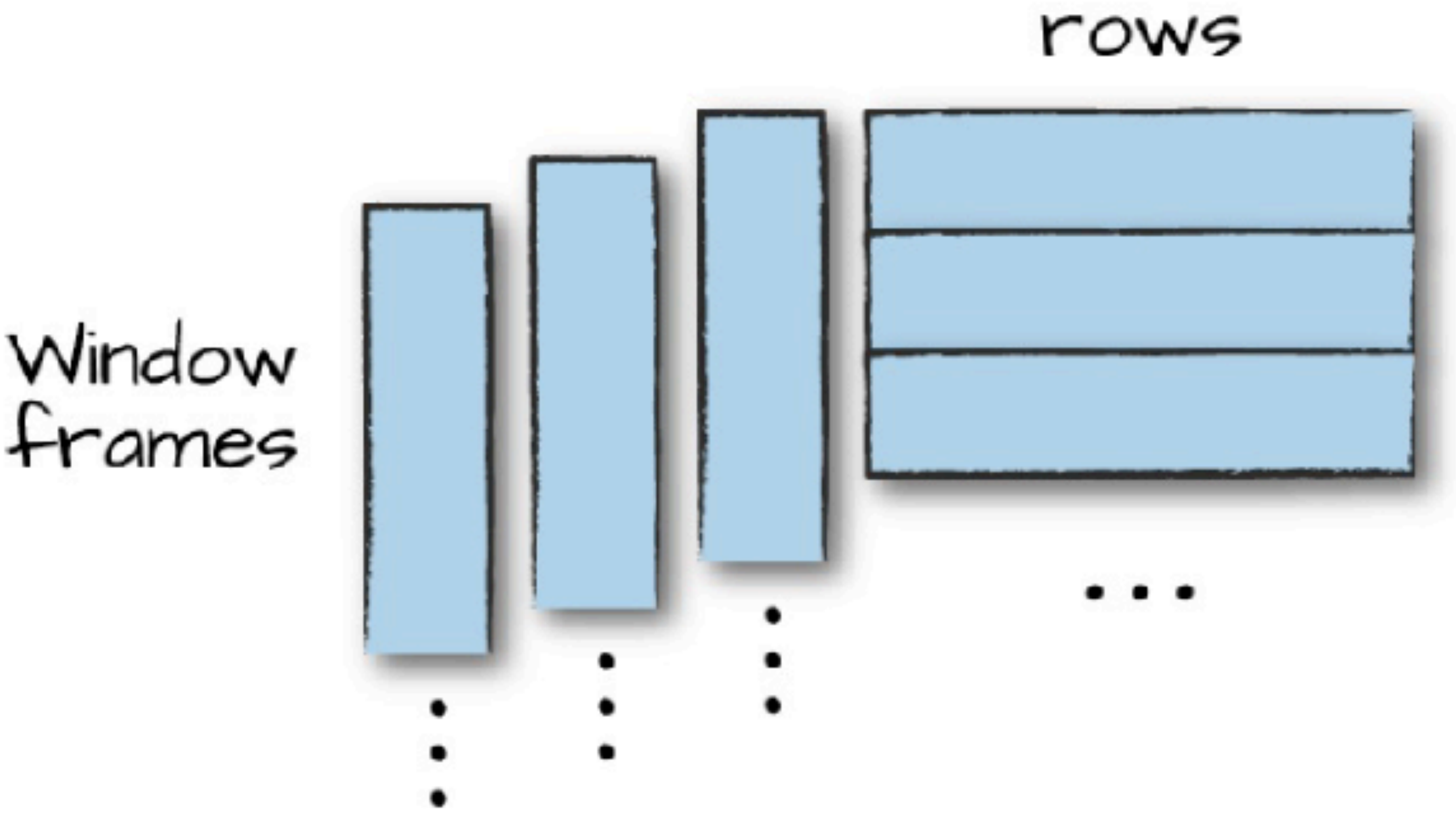


Figure 7-1. Visualizing window functions

Source: Chamber, Bill et al. 2017. Spark: The Definitive Guide, pg 135

II

Prepping Data For Processing

MITGOV/LAB

CIVIC DESIGN DATA LAB

DIRECTORATE OF SCIENCE TECHNOLOGY & INNOVATION

africell

II Prepping Data for Processing

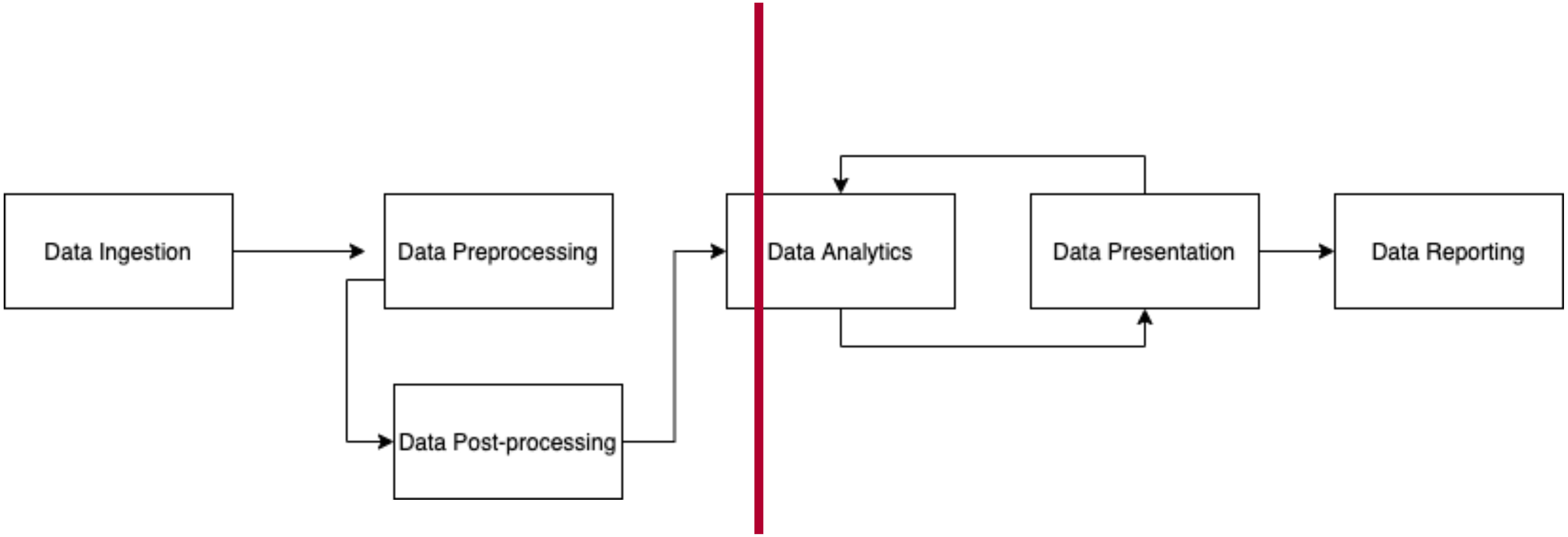


Figure 1: *Basic Big Data Analytics Processing*

II Prepping Data for Processing

Our analyses require other types of data:

1. **shapefiles** for districts, chiefdoms, sections
2. antenna and site **meta-data**

We need to be intentional in our processing of this data. The decision we make here will impact the rest of our analysis. Time should be spent cleaning and understanding these datasets.

II Prepping Data for Processing

Follow on Zeppelin

III

Creating Aggregates

MITGOV/LAB

CIVIC DESIGN DATA LAB

DIRECTORATE OF SCIENCE TECHNOLOGY & INNOVATION

africell

III Creating Aggregates

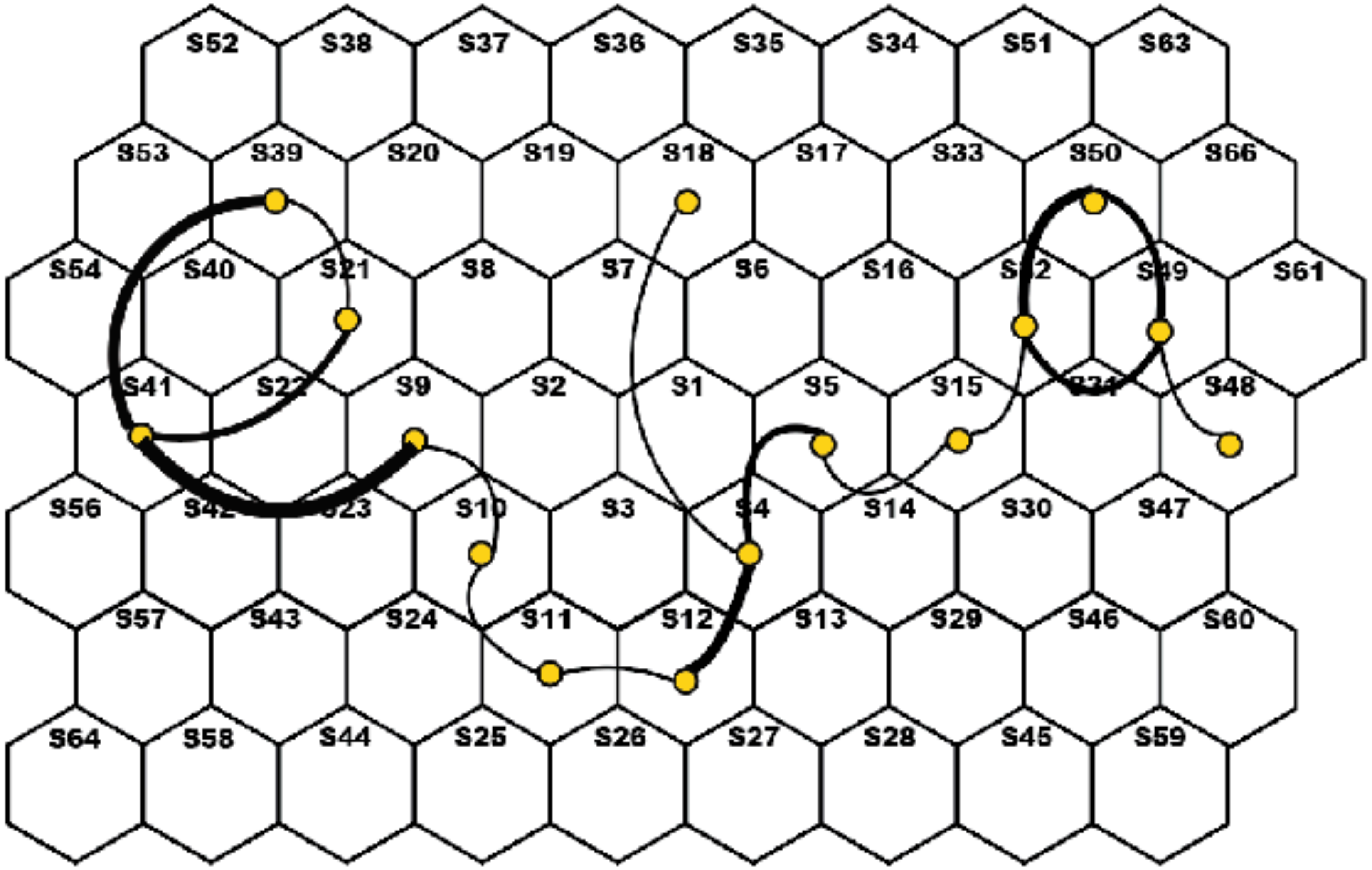
We will review to types of aggregation task in Spark:

1. **custom aggregations**
    1. generating stops and journeys
2. **off-the-shelf aggregations**
    1. running Flowminder SQL Queries

III Creating Aggregates: stops & journeys

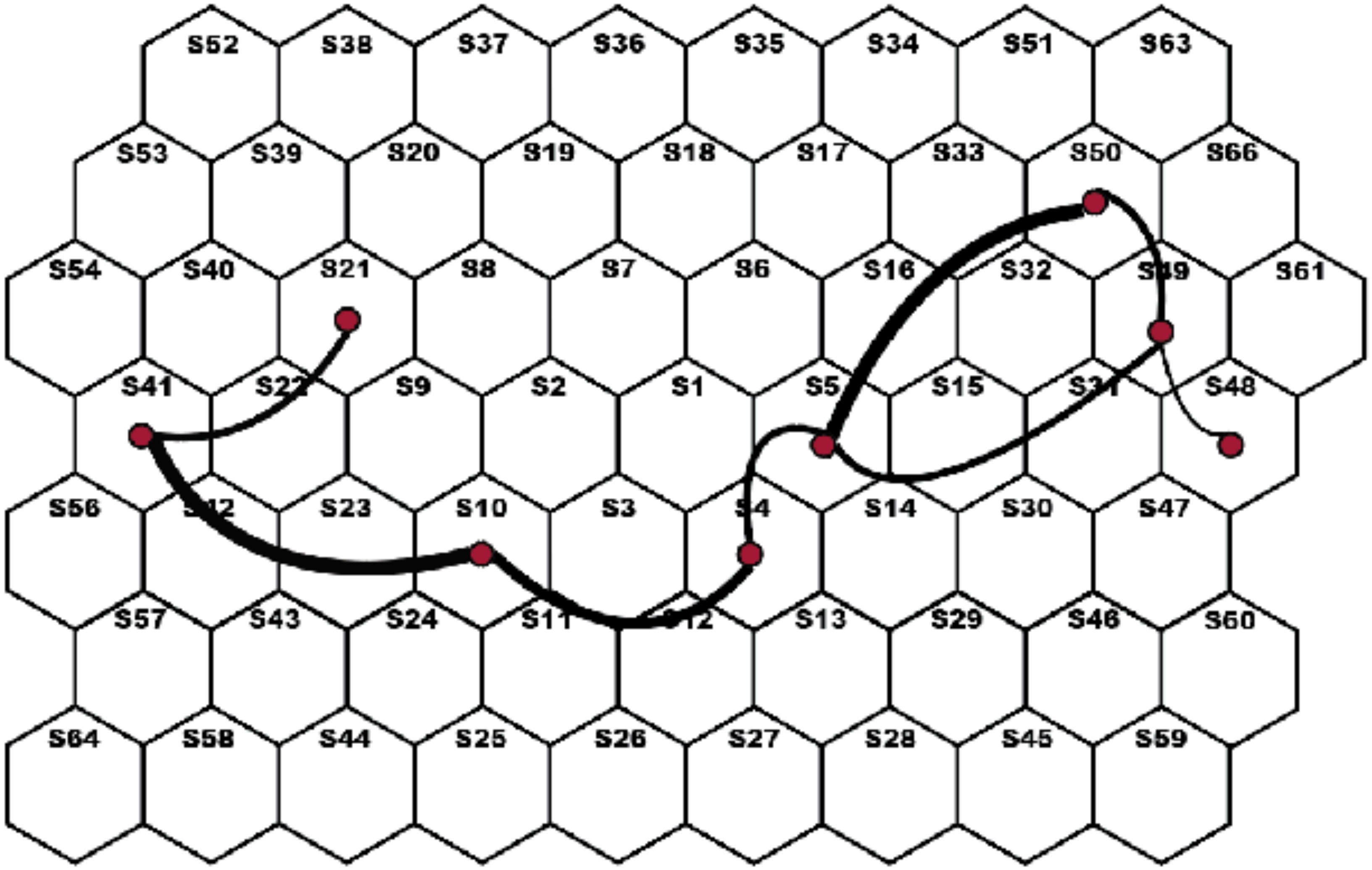
subscriber QWD1823HD81238NM1

Movement on 02/29/2020



III Creating Aggregates: stops & journeys

subscriber QWD1823HD81238NM1
Movement on 02/29/2020



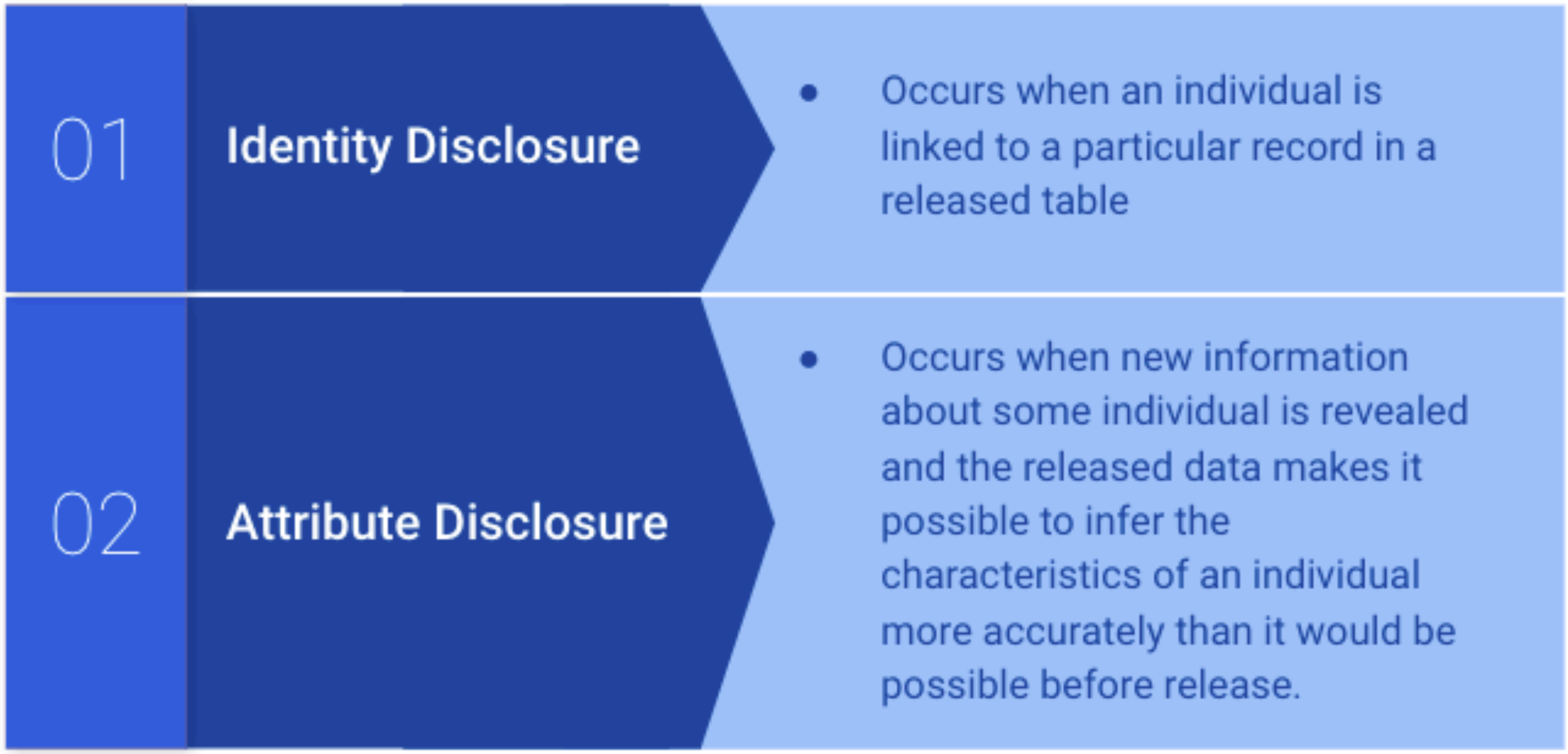
III Creating Aggregates

Follow on Zeppelin

IV

Closing

IV Closing: Privacy

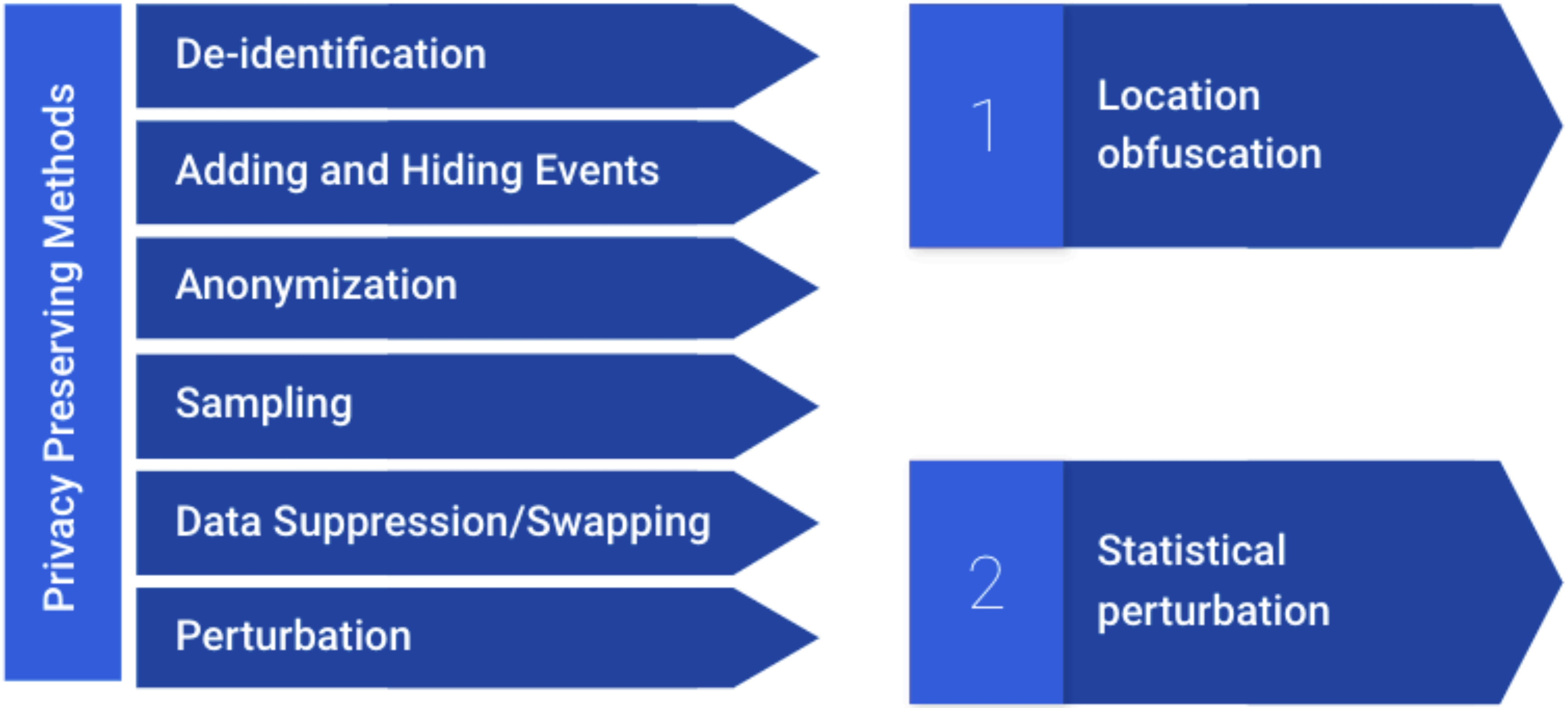


01	Identity Disclosure	- Occurs when an individual is linked to a particular record in a released table
02	Attribute Disclosure	- Occurs when new information about some individual is revealed and the released data makes it possible to infer the characteristics of an individual more accurately than it would be possible before release.

For a comprehensive overview, see *Protecting User Data and Privacy*

MITGOV/LAB CIVIC DESIGN DATA LAB DIRECTORATE OF SCIENCE TECHNOLOGY & INNOVATION africell

IV Closing: Privacy



Privacy Preserving Methods

- De-identification
- Adding and Hiding Events
- Anonymization
- Sampling
- Data Suppression/Swapping
- Perturbation

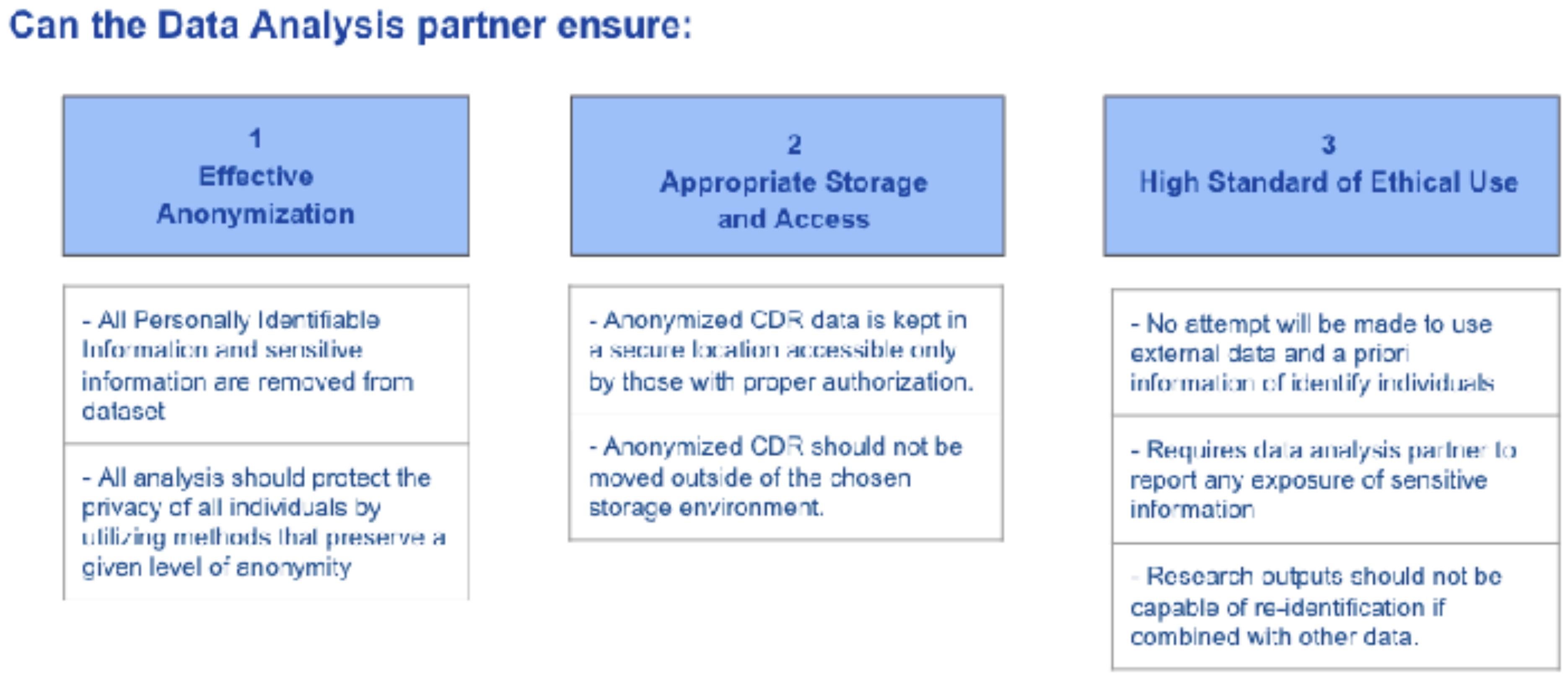
1 Location obfuscation

2 Statistical perturbation

For a comprehensive overview, see *Protecting User Data and Privacy*

MITGOV/LAB CIVIC DATA DESIGN LAB DIRECTORATE OF SCIENCE TECHNOLOGY & INNOVATION africell

IV Closing: Privacy



Can the Data Analysis partner ensure:

1 Effective Anonymization	2 Appropriate Storage and Access	3 High Standard of Ethical Use
- All Personally Identifiable Information and sensitive information are removed from dataset	- Anonymized CDR data is kept in a secure location accessible only by those with proper authorization.	- No attempt will be made to use external data and a priori information of identify individuals
- All analysis should protect the privacy of all individuals by utilizing methods that preserve a given level of anonymity	- Anonymized CDR should not be moved outside of the chosen storage environment.	- Requires data analysis partner to report any exposure of sensitive information
		- Research outputs should not be capable of re-identification if combined with other data.

For a comprehensive overview, see *Protecting User Data and Privacy*

MITGOV/LAB

CIVIC DESIGN DATA LAB

DIRECTORATE OF SCIENCE TECHNOLOGY & INNOVATION

africell

IV Closing: Much More

- CDR can be use to model social networks. Spatial and temporal interactions among subscribers

- Can be overplayed with other data: census, business, survey.

- Apache Spark can handle this! Support for both streaming and batch analytics. Integrations pipelines for large-scale Machine Learning.

MITGOV/LAB CIVIC DESIGN DATA LAB DIRECTORATE OF SCIENCE TECHNOLOGY & INNOVATION africell

Thank You!

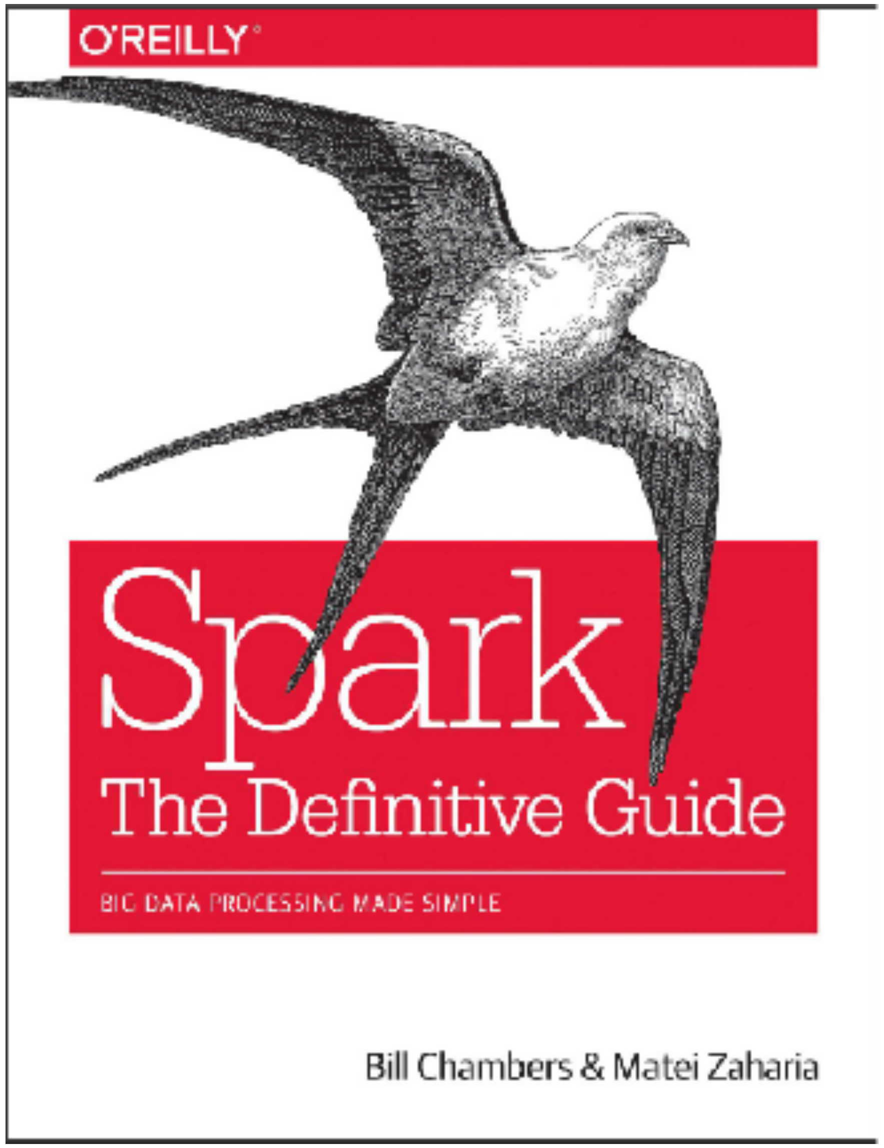
MITGOV/LAB

CIVIC DESIGN
DATA LAB

DIRECTORATE OF SCIENCE
TECHNOLOGY & INNOVATION

africell

AX References



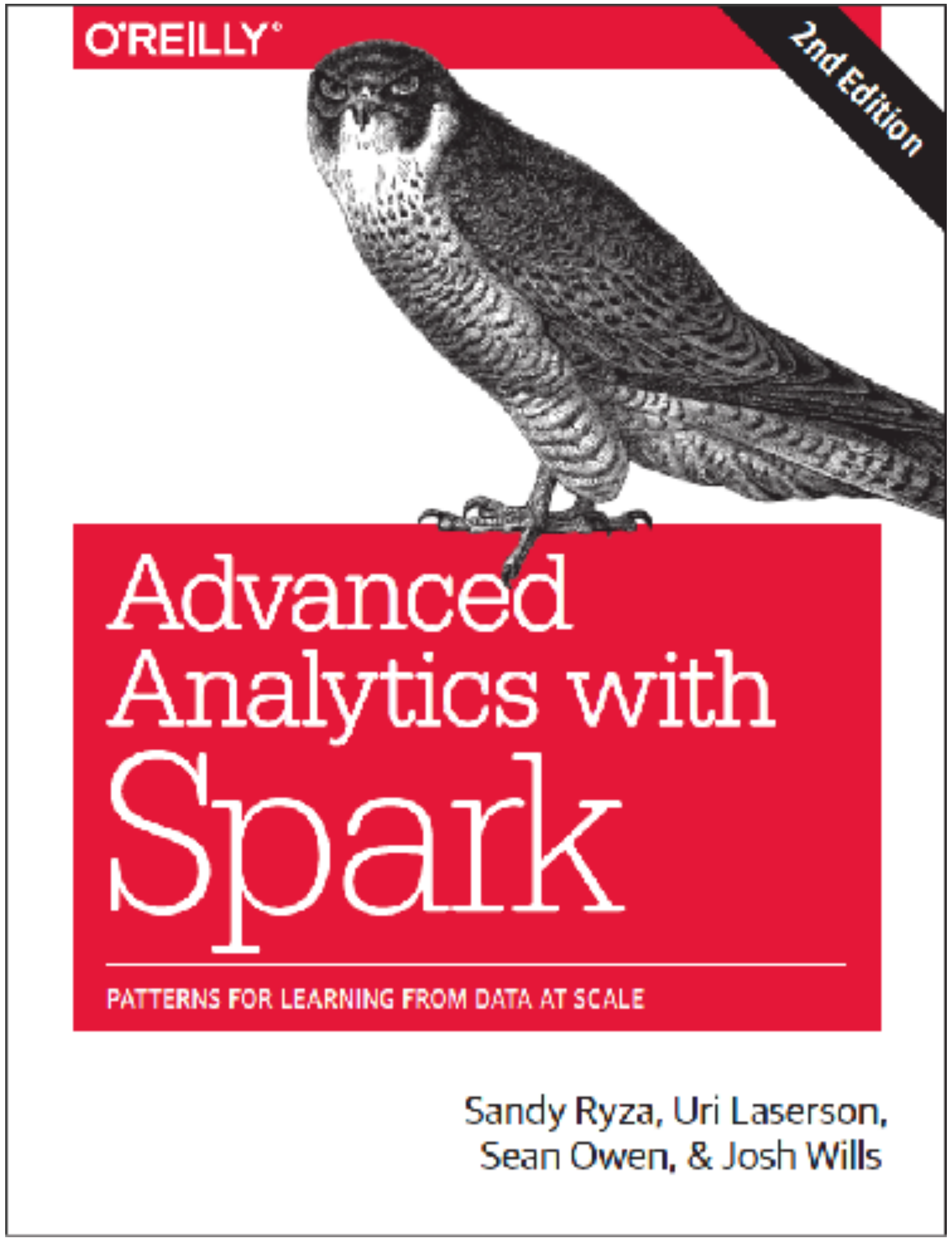
O'REILLY

Spark

The Definitive Guide

BIG DATA PROCESSING MADE SIMPLE

Bill Chambers & Matei Zaharia



O'REILLY

2nd Edition

Advanced Analytics with Spark

PATTERNS FOR LEARNING FROM DATA AT SCALE

Sandy Ryza, Uri Laserson, Sean Owen, & Josh Wills

MITGOV/LAB

CIVIC DESIGN DATA LAB

DIRECTORATE OF SCIENCE TECHNOLOGY & INNOVATION

africell