

The background is a dark blue gradient with a subtle pattern of white dots. On the left side, there are several overlapping circular elements. One prominent circle has a scale with numbers ranging from 140 to 260. Other circles contain dashed lines and arrows, suggesting a process or cycle. The overall aesthetic is technical and modern.

ACQUIRING AND PREPARING DATA

TOPICS

- Content Acquisition Approaches, Pros & Cons
- Working with Beautiful Soup
- Acquiring data using Rest Based APIs
- Data Cleaning & Wrangling using Pandas
- Missing Values and Outlier

//

Data acquisition is the processes for bringing data that has been created by a source outside the organization, into the organization, for production use.

<https://www.firstsanfranciscopartners.com/blog/defining-data-acquisition-importance/?cn-reloaded=1>

REAL TIME

Data should be processed in a real-time manner without any time delay. Normally, there would be a demand for real-time processing in trading-related contexts. For example:

- For online trade fraud prevention, the data of trading parties should be dealt with by an anti-fraud model at the fastest possible speed, to judge if there is any fraud, and promptly report any deviant behavior to the authorities.
- The commodities of an ecommerce website should be recommended in a real-time manner according to the historical data of clients and the current web page browsing behavior.
- Computer manufacturers should, according to their sales conditions, make a real-time adjustment of inventories, production plans, and parts supply orders.
- The manufacturing industry should, based on sensor data, make a real-time judgment of production line risks, promptly conduct troubleshooting, and guarantee the production.

MICRO BATCH

Data should be processed by the minute in a periodic manner. It is not necessary that data is processed in a real-time manner. Some delay is allowed. For example, the effect of an advertisement should be monitored every five minutes to determine a future release strategy. It is thus required that data should be processed in a centralized manner every five minutes in aggregate.

MEGA BATCH

Data should be processed periodically with a time span of several hours, without a high volume of data ingested in real time and a long delay in processing. For example, some web pages are not frequently updated and web page content may be crawled and updated once every day.

STREAMING DATA

Streaming data is not necessarily acquired in a real-time manner. It may also be acquired in batches, depending on application context. For example, the click event stream of a mobile app is uploaded in a continuous way. However, if we only wish to count the added or retained stream in the current day, we only need to incorporate all click-stream logs in that day in a document and upload them to the system by means of a mega batch for analytics.

DATA INGESTION

- Data ingestion refers to a process by which the data acquired from data sources is brought into your system, so the system can start acting upon it. It concerns how to acquire data.
- Data ingestion typically involves three operations, namely discover, connect, and sync. Generally, no revision of any form is made to numeric values to avoid information loss.

DATA INGESTION OPERATIONS

- **Discover** refers to a process by which accessible data sources are searched in the corporate environment. Active scanning, connection, and metadata ingestion help to develop the automation of the process and reduce the workload of data ingestion.
- **Connect** refers to a process by which the data sources that are confirmed to exist are connected. Once connected, the system may directly access data from a data source. For example, building a connection to a MySQL database involves configuring the connecting strings of the data source, including IP address, username and password, database name, and so on.
- **Sync** refers to a process by which data is copied to a controllable system. Sync is not always necessary upon the completion of connection. For example, in an environment which requires highly sensitive data security, only connection is allowed for certain data sources. Copying is not allowed for that data.

BEAUTIFUL SOUP

Beautiful Soup is a Python library for pulling data out of HTML and XML files. It works with your favorite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work.

- <https://beautiful-soup-4.readthedocs.io/en/latest/>

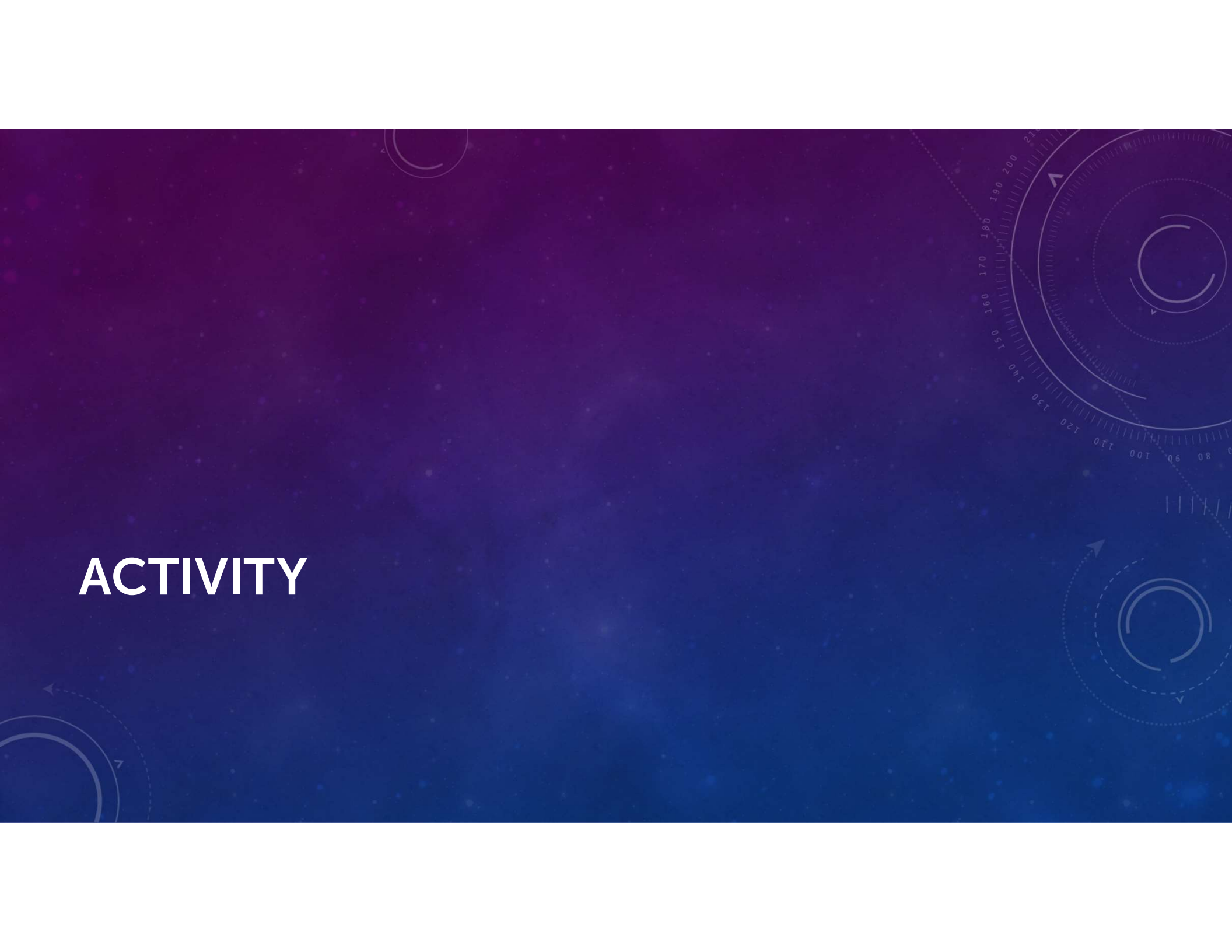
ACQUIRING DATA USING REST BASED APIS

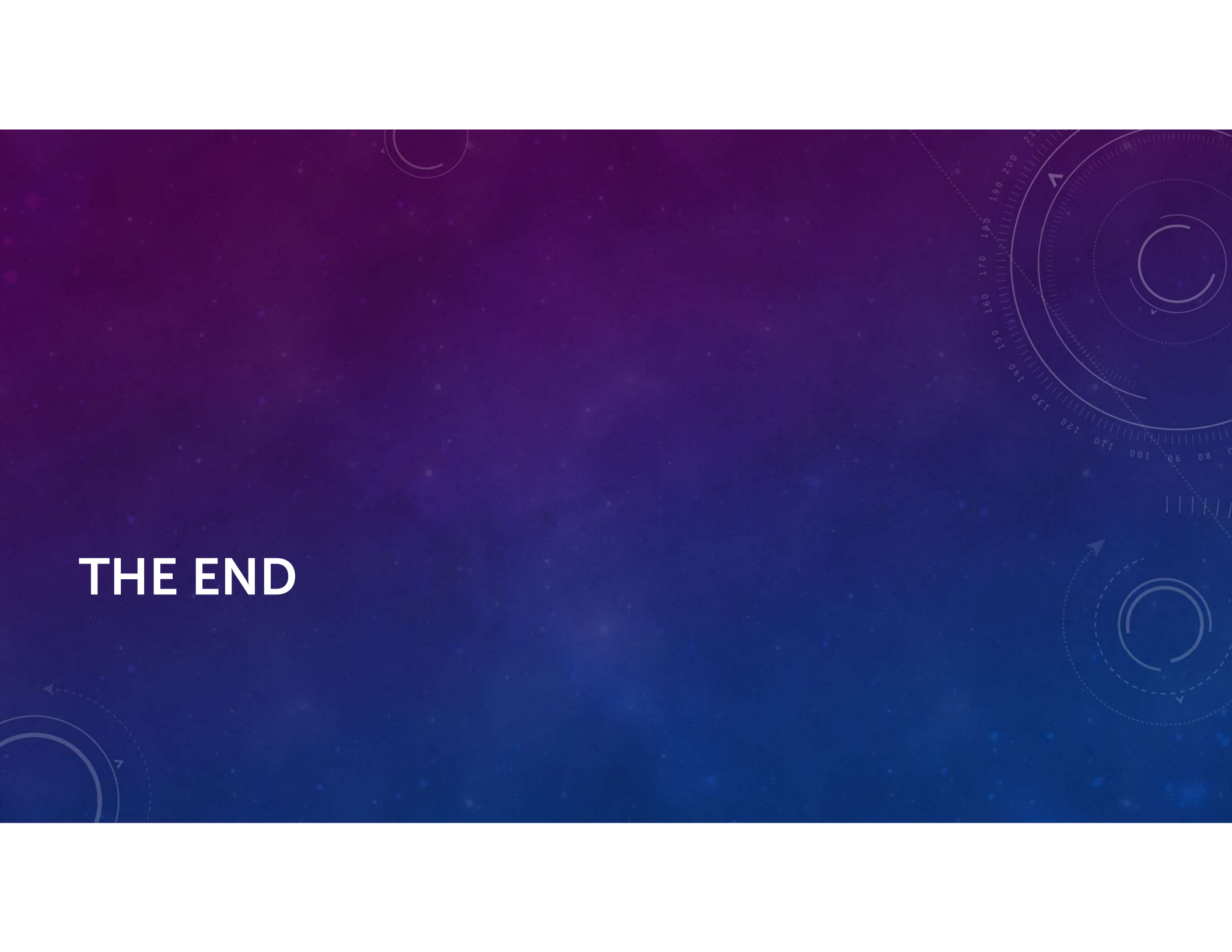
The background is a dark blue gradient with faint, light blue circular patterns and a large circular scale on the right side. The scale has numbers from 0 to 210 and a dashed line with an arrow pointing clockwise. There are also some faint, light blue dots scattered across the background.

DATA CLEANING & WRANGLING USING PANDAS

The background is a dark blue gradient with a starry space theme. It features several faint, light blue circular patterns and arcs, some of which resemble clock faces or data visualizations. A prominent circular scale with numerical markings (0, 90, 180, 270) is visible in the upper right corner. The overall aesthetic is technical and modern.

ACTIVITY

The background is a dark blue gradient with a subtle pattern of white stars. Faint, light blue circular patterns with arrows are visible in the corners, suggesting a technical or scientific theme.



THE END