Steam videogames platform

Exploratory Data Analysis

Context

- Conduct a study of the videogames market

- Data from Steam videogames platform’s catalogue

  Release date, number of owners, genre, langages, ...

- Assume a large dataset that requires distributed computation

Context

- Conduct a study of the videogames market
- Data from Steam videogames platform’s catalogue

  Release date, number of owners, genre, langages, …
- Assume a large dataset that requires distributed computation

Tasks

- Load the dataset in a Spark context
- Preprocess the dataset into tables relevant for data analysis
- Carry out the exploratory data analysis using PySpark API
- It is proposed to use cloud services such as Databricks to conduct the EDA and make visualizations

Data processing

- Data loading in Spark
  - Direct download from AWS S3
  - Download file locally (61Mb)

Data processing

Data loading in Spark

- Direct download from AWS S3
- Download file locally (61Mb)

Data processing in PySpark

- DataFrame API
  Process with object methods
- SQL API
  Process with SQL queries
- Pandas API
  Convert to `pandas.DataFrame`

Data processing

Data loading in Spark

- Direct download from AWS S3
- Download file locally (61Mb)

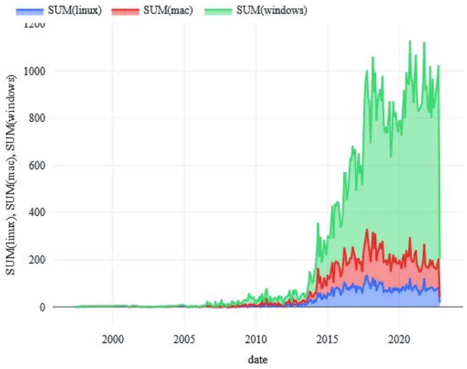
Data processing in PySpark

- DataFrame API
  Process with object methods
- SQL API
  Process with SQL queries
- Pandas API
  Convert to `pandas.DataFrame`

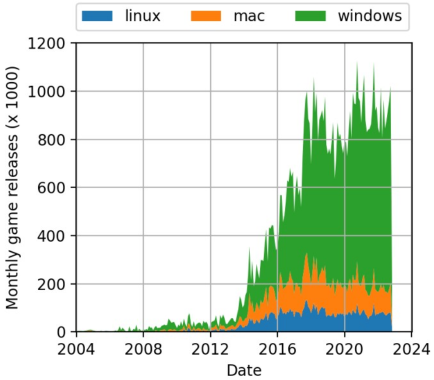
Data visualization

- Export to Python structures (`pandas.DataFrame`, `np.ndarray`, `list`) and build visualization (matplotlib, plotly, ...)
- Use vizualization tools from cloud services (eg Databricks)

Databricks plotly figure



Matplotlib figure



Comparison of the technologies used

pandas

- ✔ Python-like interface
- ✔ Very flexible
  full Python expressivity
- ✘ Non-distributed compute
  Inefficient for large datasets
- ✘ Requires Python

PySpark

- ✔ Distributed computations
- ✔ Flexible
- ✘ Slow computation setup
  Inefficient for small datasets
- ✘ Limited to a few langages
  Java/Scala/Python/R

SQL

- ✔ Engine agnostic
  Compatible with many frameworks
- ✔ Distributed computations
- ✘ Not flexible
  Difficult to carry complex data transformations

Comparison of the technologies used

pandas

- ✔ Python-like interface
- ✔ Very flexible
  full Python expressivity
- ✘ Non-distributed compute
  Inefficient for large datasets
- ✘ Requires Python

PySpark

- ✔ Distributed computations
- ✔ Flexible
- ✘ Slow computation setup
  Inefficient for small datasets
- ✘ Limited to a few langages
  Java/Scala/Python/R

SQL

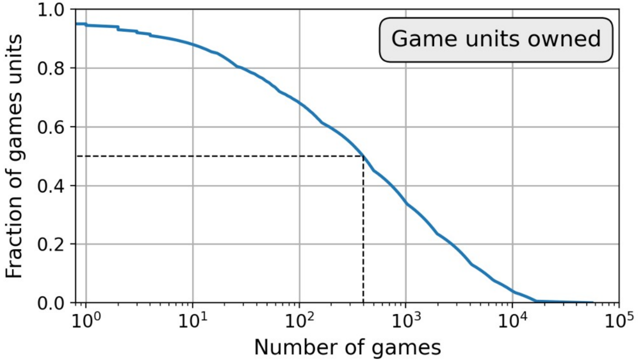
- ✔ Engine agnostic
  Compatible with many frameworks
- ✔ Distributed computations
- ✘ Not flexible
  Difficult to carry complex data transformations

Data processing strategy

- Large-scale operations with PySpark
- Small-scale processing with Python

Key takeaways on the videogames market

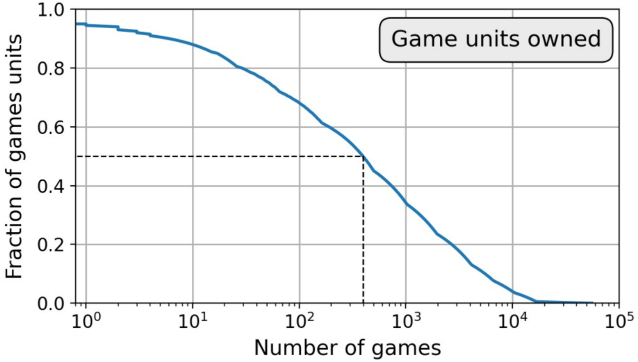
- Videogames market dominated by a few big companies
  - Half units distrubuted from 400 games

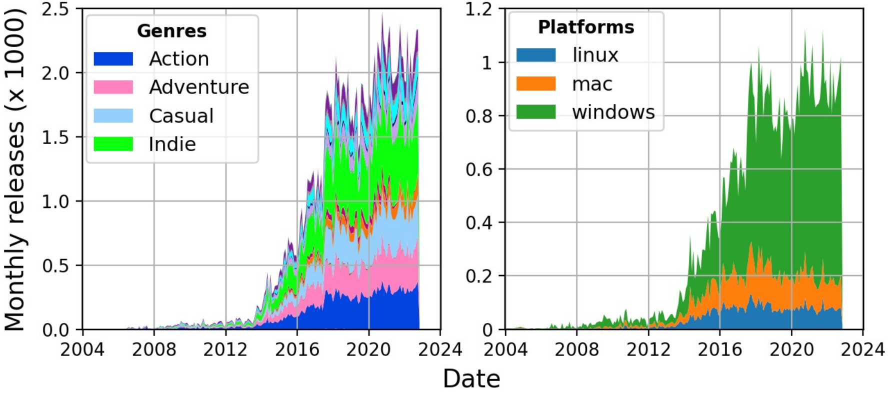


Key takeaways on the videogames market

- Videogames market dominated by a few big companies
  - Half units distrubuted from 400 games

- Game release increase until 2018 then stabilized
  - Windows is the dominant platform
  - Major game genres : action, adventure, causal and indie



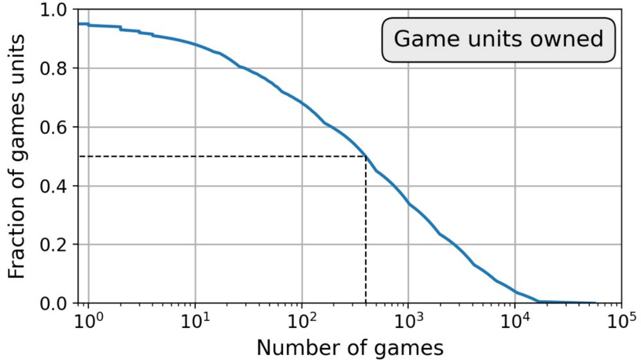


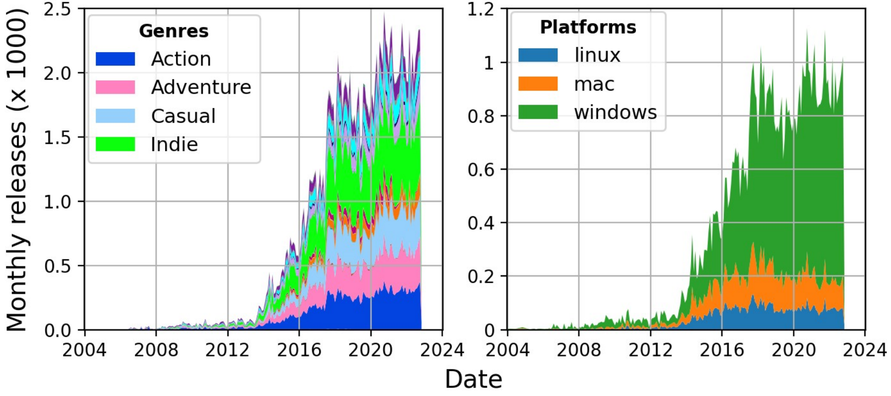
Key takeaways on the videogames market

- Videogames market dominated by a few big companies
  - Half units distrubuted from 400 games

- Game release increase until 2018 then stabilized
  - Windows is the dominant platform
  - Major game genres : action, adventure, causal and indie

- Microtransactions are an important source of revenue
  - Major free games : Dota 2, ...





Thanks!