# Data Engineering for the Masses
## Going from IT-Experts to Everyday Users

Data Engineering and Science are currently highly specialized areas that require deep knowledge in their respective domains. We present four major trends and needs that are necessary to open these areas to non-specialists in the future.

```
> ./start-spark.sh
  Spark running...
```

Start Spark ✓
Stop Spark ✗

Moving from CMD scripts to GUIs.
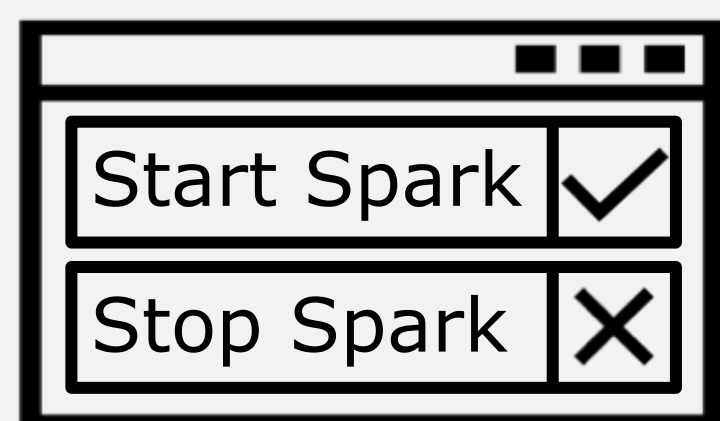
Most non-IT people have never seen or used the command line before. If data engineering tasks are only usable from there, people won't use it. Providing graphical interfaces opens the realm for less technical users.

Spark AWS Hadoop

Apache what?
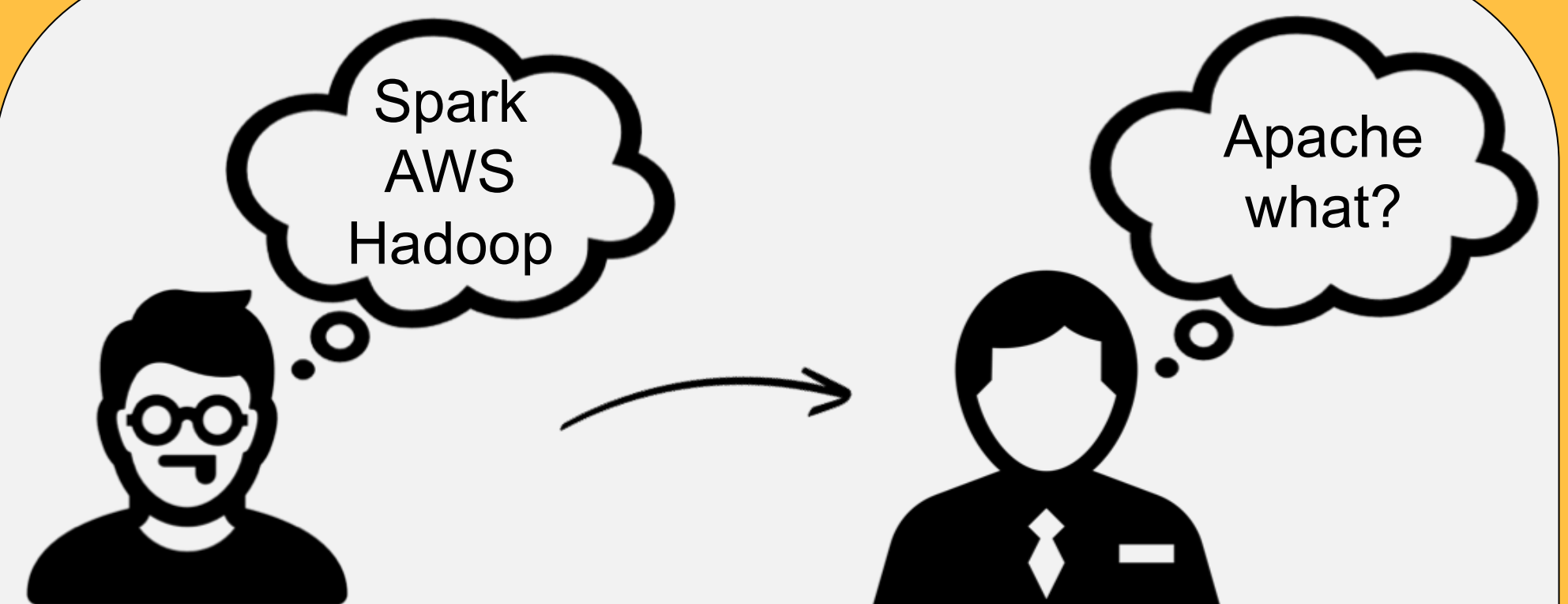
Moving from specialized frame-works, databases, and systems to technology-stack agnosticism.

Users rarely care about the specific technology that is used. Providing powerful abstractions on top of the underlying stack, e.g. Apache Spark, Hadoop, or Amazon AWS will allow more non-specialists to use these systems.

Normalise → **Clean** → **Dedup**

Preparing your data...

Moving from complicated ETL jobs to automated data upload.

Many preprocessing tasks, e.g. data normalisation, cleansing, or duplicate removal require specialist knowledge. Fully automating these jobs (with minimal user interaction) will not only reduce the time before the actual analysis but also allow non-specialists to work with higher quality data.

```
0.5,1,a,136
0.6,3,b,254
0.9,9,a,442
```

Moving from CSV exports to visual results.

Many analytics tools export their data in raw formats that can only be interpreted with further tools. To allow non-experts to directly interpret the results, good visualisation needs to be integrated into the pipeline.

Lawrence Benson
in M.Sc. IT-Systems Engineering

Hasso Plattner Institute, Potsdam, Germany

E-Mail: lawrence.benson@student.hpi.de

Referenced Lectures

Visualisation	: Tableau
Tech-Stack	: Amazon Redshift
GUIs	: bakdata

HPI Hasso Plattner Institut