Big Data Tools:

Holy Grail of Deveoper Productivity

Pasha Finkelshteyn, JetBrains

Who am I

- ex system administrator
- ex developer
- ex team lead
- ex data egineer
- developer advocate for big data

Together >14 years in IT

asm0di0 BigDataToolsJB

Who are data engineers?

Responsibilities:

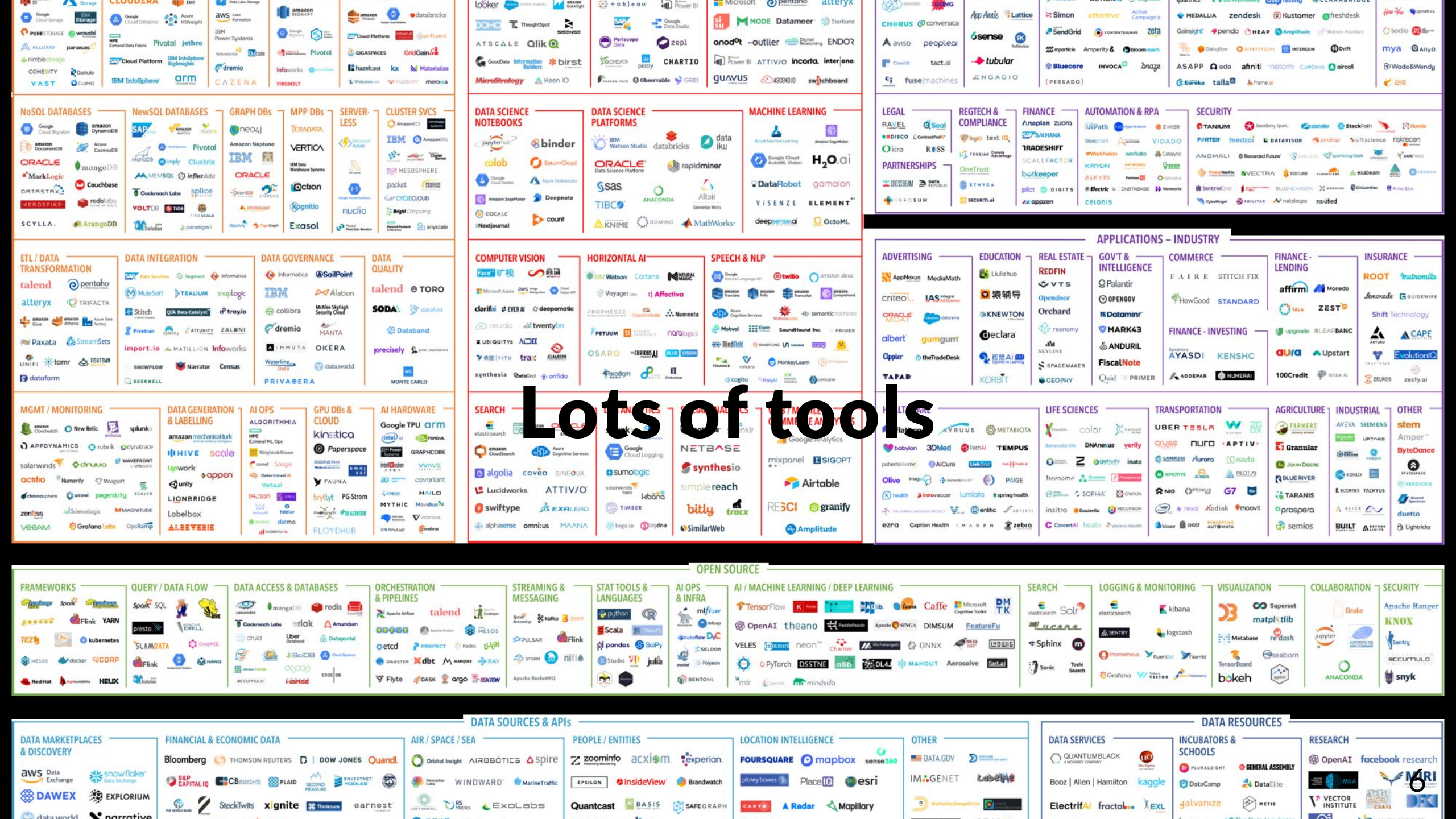
- Build your DWH
- Build your DMP
- Transfer and store your data

As effective and fast as possible

I know pain points of data engineering

I know pain points of data engineering

And today I'll try to solve them for you


Lots of tools
NoSQL DATABASES
NewSQL DATABASES
GRAPH DBs
MPP DBs
SERVERLESS
CLUSTER SVCS
DATA SCIENCE NOTEBOOKS
DATA SCIENCE PLATFORMS
MACHINE LEARNING
LEGAL
REGTECH & COMPLIANCE
FINANCE
AUTOMATION & RPA
SECURITY
PARTNERSHIPS
ETL / DATA TRANSFORMATION
DATA INTEGRATION
DATA GOVERNANCE
DATA QUALITY
COMPUTER VISION
HORIZONTAL AI
SPEECH & NLP
APPLICATIONS – INDUSTRY
ADVERTISING
EDUCATION
REAL ESTATE
GOV'T & INTELLIGENCE
COMMERCE
FINANCE - LENDING
INSURANCE
FINANCE - INVESTING
MGMT / MONITORING
DATA GENERATION & LABELLING
AI OPS
GPU DBs & CLOUD
AI HARDWARE
SEARCH
LIFE SCIENCES
TRANSPORTATION
AGRICULTURE
INDUSTRIAL
OTHER
OPEN SOURCE
FRAMEWORKS
QUERY / DATA FLOW
DATA ACCESS & DATABASES
ORCHESTRATION & PIPELINES
STREAMING & MESSAGING
STAT TOOLS & LANGUAGES
AI OPS & INFRA
AI / MACHINE LEARNING / DEEP LEARNING
SEARCH
LOGGING & MONITORING
VISUALIZATION
COLLABORATION
SECURITY
DATA SOURCES & APIs
DATA MARKETPLACES & DISCOVERY
FINANCIAL & ECONOMIC DATA
AIR / SPACE / SEA
PEOPLE / ENTITIES
LOCATION INTELLIGENCE
OTHER
DATA RESOURCES
DATA SERVICES
INCUBATORS & SCHOOLS
RESEARCH

Lots of tools

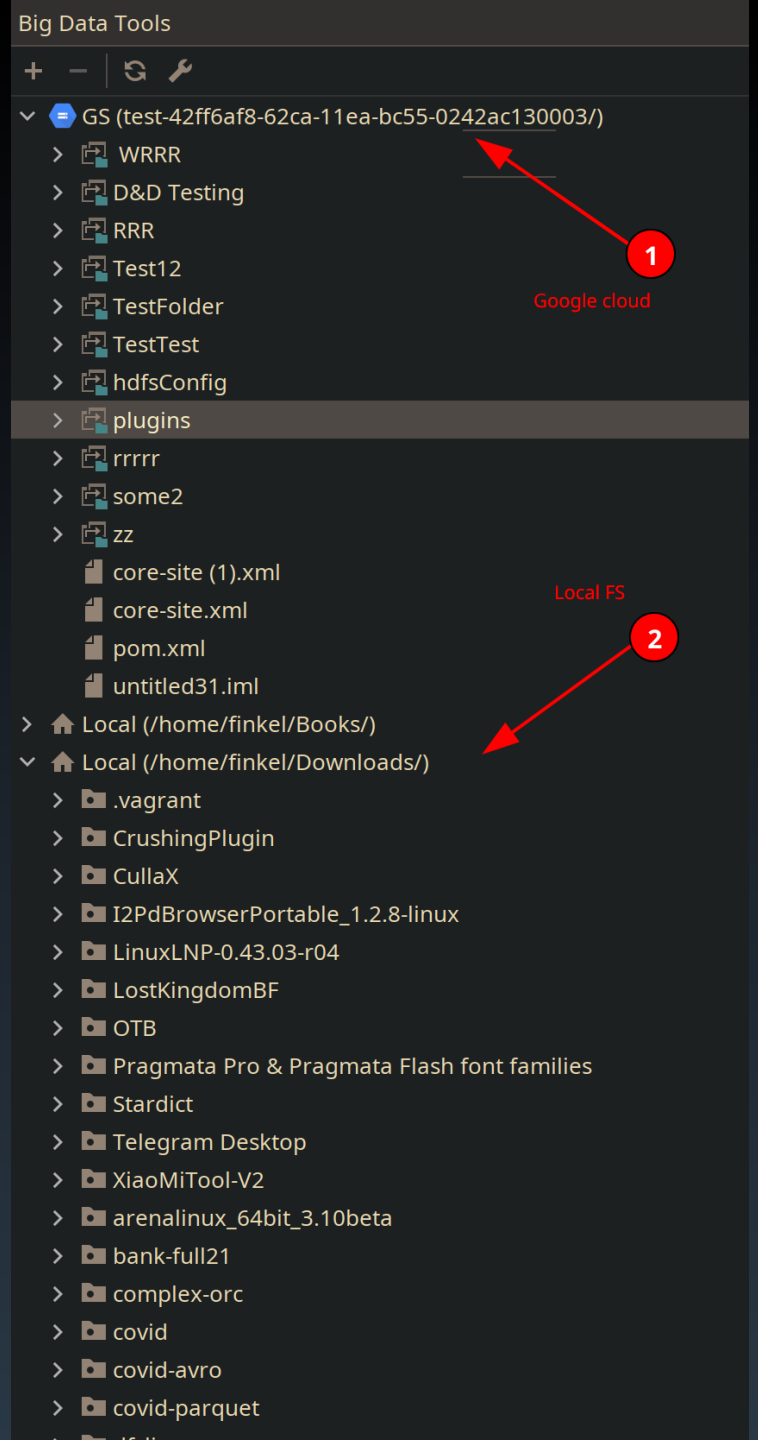
- Nobody may know everything
- Constant context switches
- No single point of work

Ultimate Big Data Workspace

Single IDE to work with Kafka, Spark, and storages of different types

Remote FS support

- Drag'n'Drop support
- Local ↔ Remote
- Remote ↔ Remote
- Basic operations:
  - Rename
  - Move
  - Copy
  - Delete



Kafka support

- Consumers (and groups)
- Topic metadata
  - Replicas
  - Partitions (and info on them)
  - Under replicated partitions

Apache Zeppelin support

> Web-based notebook that enables data-driven, interactive data analytics and collaborative documents with SQL, Scala and more.

Zeppelin is awesome environment to run ad-hoc tasks.

But this is not the best IDE humanity developed.

Create Dataset/DataFrame via SparkSession

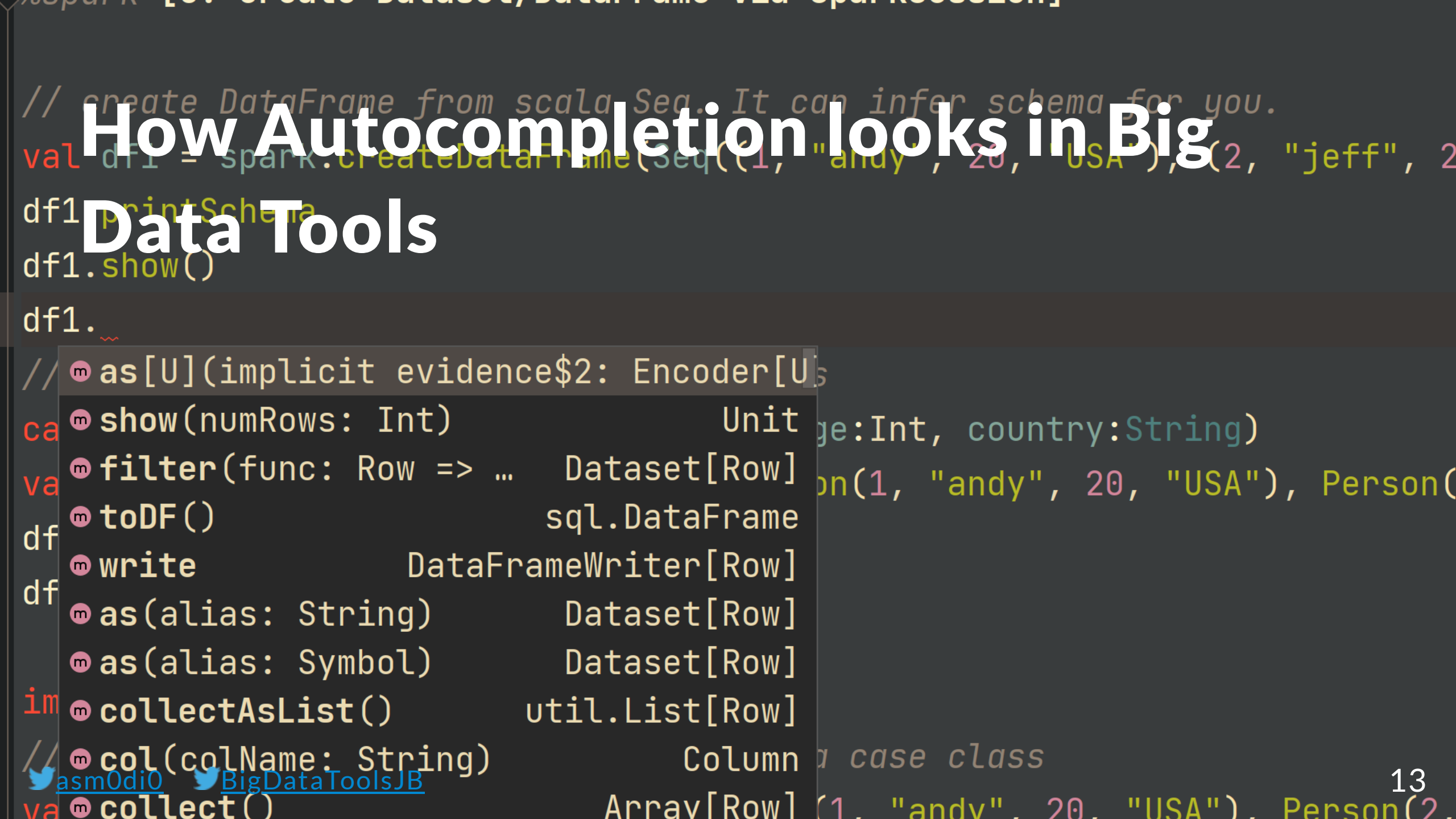
How autocompletion looks in web version ovf Zeppelin

```
%spark
// create DataFrame from scala Seq. It can infer schema for you.
val df1 = spark.createDataFrame(Seq((1, "andy", 20, "USA"), (2, "jeff", 23, "China"), (3, "james", 18, "USA"
df1.printSchema
df1.show()
df1.
// c
case
val                                                    Person(2, "jeff", 23, "China"), Person(3,
df2.
df2.

impo
// y
val                                                    erson(2, "jeff", 23, "China"), Person(3, "j
df3.
df3.
```

```
abc China
abc DataFrame
abc Dataset
abc Int
abc It
abc Person
abc Seq
abc String
abc USA
abc _
abc age
abc also
```

How Autocompletion looks in Big Data Tools



Autocompletion

Autocompletion is one of the key factors of developer productivity today.

IntelliJ IDEA uptime: 12 sec, 809 ms, idle time: 134 ms
Code completion has saved you from typing at least 117.6K characters since 11.02.2020 (~497 per working day)
Quick fixes have saved you from 5,720 possible bugs since 11.02.2020 (~25 per working day)

Feature	Group	Used	Last Used
Basic code completion	Code Completion	26,274 times	one week ago
Browse external documentation	Code Assistants	Never	N/A
Camel prefixes in code completion	Code Completion	2,464 times	one week ago
Cancel lookup and move caret up/...	Code Completion	Never	N/A
Changing completion variants sorting	Code Completion	Never	N/A
Comment/Uncomment block	Code Assistants	114 times	2 months ago
Comment/Uncomment current line	Code Assistants	761 times	one week ago

Code Completion **Ctrl+Space** helps you quickly complete code statements. It works as you type and gives a list of suggestions available from the current caret position:

```
try {
    service = Executors.newSingleThreadExecutor();
    f = service.s
} finally {
    if (servi
    System.ou
}
while (!f.isD
    try {
```

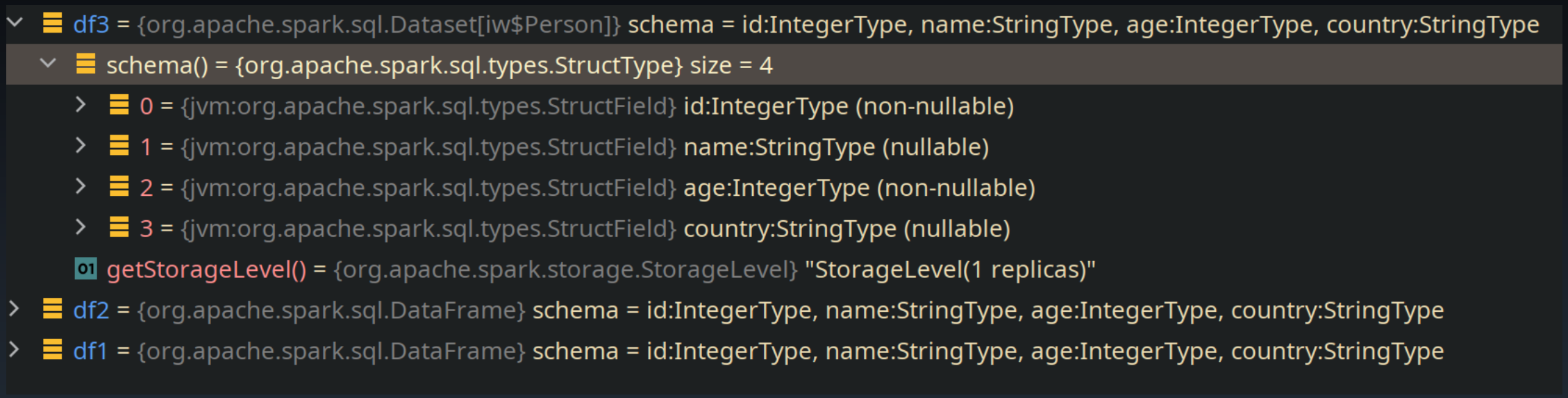
```
submit(Runnable task)                Future<?>
submit(Callable<T> task)             Future<T>
submit(Runnable task, T result)      Future<T>
shutdown()                           void
Press ⏎ to insert, ⇥ to replace
```

IDE features help!

We're saving insane amount of time: preventing bugs, saving from typing, adding imports automatically...

Features specific for Data Engineers

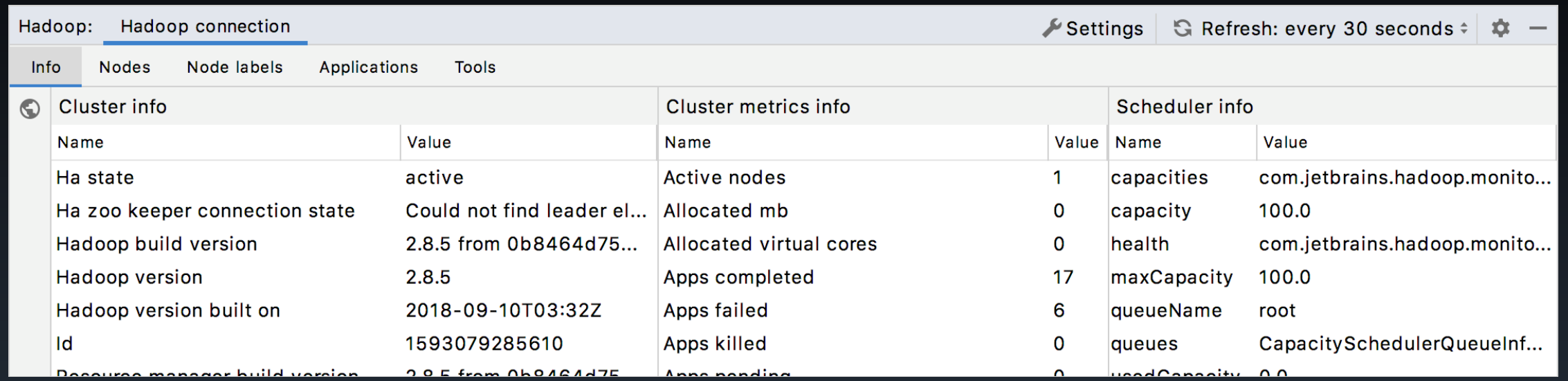
Structure of dataframe



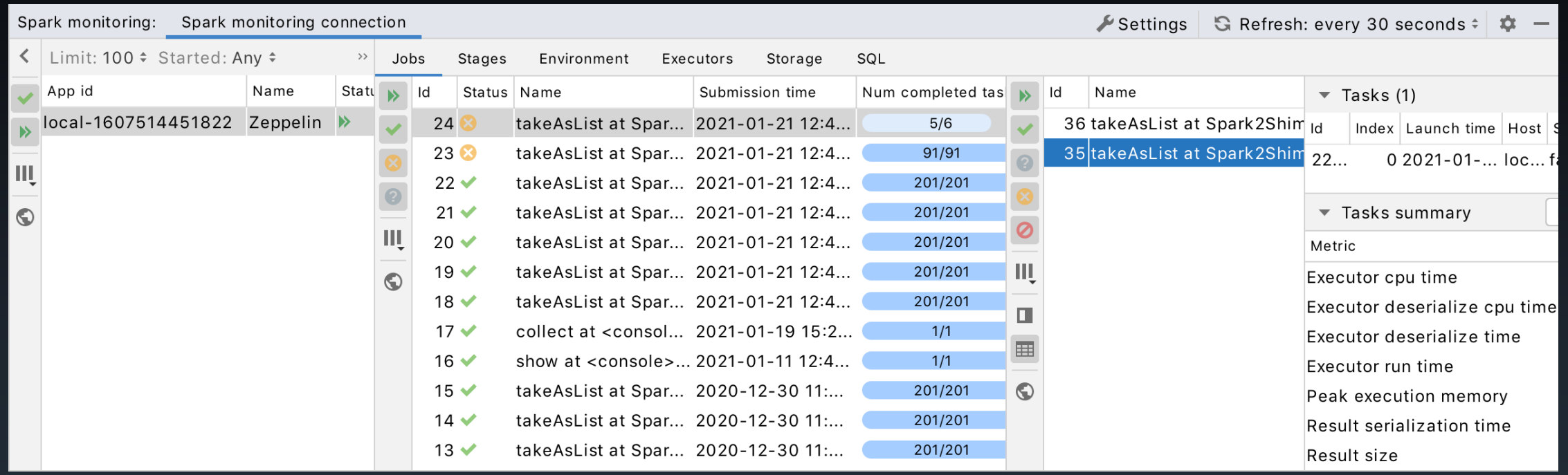
This information is not available in any environment but Big Data Tools

The same applies to Python

Hadoop Monitoring



Spark Monitoring



Conclusion

Conclusion

- Smart features
  - Autocompletion
  - Data introspection
- Monitoring
  - Hadoop
  - Spark
- Work with popular Object Storages
- Single point of integration

Big Data Tools plugin

Thank you!

Pasha Finkelshteyn, JetBrains

asm0di0 BigDataToolsJB

Welcome to Q&A