

SCALA , REDIS and PowerBI

Research Slides

Preeti Sehgal



Scala



What is SCALA:

- Scala, short for Scalable Language, is a hybrid functional programming language.
- It was created by Martin Odersky.
- Scala smoothly integrates the features of object-oriented and functional languages.
- Scala is compiled to run on the Java Virtual Machine. Many existing companies, who depend on Java for business critical applications, are turning to Scala to boost their development productivity, applications scalability and overall reliability.



Why SCALA for Data Engineering:

- SCALA is a java based language and can be used for full blown application development based on JVM. However, for Data Engineers the real benefit of Scala is for data extraction and transformation on Spark.
- For high performance data science projects Scala is the defacto language for spark



Why did I pick Scala:

- When it comes to working on Spark; Scala and Python are the two most popular languages
- Having knowledge and experience in both the languages, Python and Scala; is a good combination for a Data Engineer



SCALA vs. PYTHON functionality and Benefits:

Topic	SCALA	PYTHON
Learning Curve	<ul style="list-style-type: none">• JAVA type programming style.• Easy for users with JAVA background• Steep learning curve for beginners with no JAVA background	<ul style="list-style-type: none">• Easy to learn• No JAVA background required
DATA Visualization	<ul style="list-style-type: none">• No good visualization tools	<ul style="list-style-type: none">• Lot of tools for data visualization
JVM	<ul style="list-style-type: none">• Requires JVM making it operating system agnostic.• This gives compatibility with Java and Interoperability	<ul style="list-style-type: none">• No JVM• All machines part of the project should be on the same OS to avoid challenges related to heterogeneous OS



SCALA vs. PYTHON functionality and Benefits: (Contd..)

Topic	SCALA	PYTHON
User Community Support	<ul style="list-style-type: none">• Smaller user community	<ul style="list-style-type: none">• Larger User community for support
Performance	<ul style="list-style-type: none">• Better performance	<ul style="list-style-type: none">• Slower performance
Streaming	<ul style="list-style-type: none">• Scala is more stable for streaming than python	<ul style="list-style-type: none">• Not as mature as Scala. When it comes to streaming, scala is preferred over python
SPARK	<ul style="list-style-type: none">• SPARK is written in Scala. So no restrictions in using Scala with Spark	<ul style="list-style-type: none">• Python works on SPARK but has some restrictions as compared to scala



SCALA vs. PYTHON Syntax Examples:

SCALA declares variables in two ways - val and var

- Val types are immutable
- Var are mutable

Reading Data from CSV File

SCALA:

```
val usersDF = spark.read.load("examples/src/main/resources/users.csv")
```

PYTHON:

```
df = spark.read.load("examples/src/main/resources/users.csv")
```

Useful Websites:

<https://spark.apache.org/docs/latest/sql-data-sources-load-save-functions.html>


<https://www.tutorialspoint.com/scala/index.htm>

Redis



What Is Redis?

- ✓ **Open source in-memory data structure store** which can be used as a database and/or a cache and message broker
- ✓ NoSQL Key/Value Store
- ✓ Supports Multiple Data Structures
- ✓ Built In Replication




Advantages Of Redis

- ✓VERY Flexible
- ✓No Schemas & Column Names
- ✓Very Fast : Can perform around **110,000 SETs per second**, about **81,000 GETs per second**
- ✓Rich Datatype Support
- ✓Caching & Disk Persistence

What Can Redis Be Used With?

ActionScript	Bash	C	C#	C++	Clojure
Common Lisp	Crystal	D	Dart	Delphi	Elixir
emacs lisp	Erlang	Fancy	gawk	GNU Prolog	Go
Haskell	Haxe	Io	Java	Julia	Lasso
Lua	Matlab	mruby	Nim	Node.js	Objective-C
OCaml	Pascal	Perl	PHP	Pure Data	Python
R	Racket	Rebol	Ruby	Rust	Scala
Scheme	Smalltalk	Swift	Tcl	VB	VCL



Redis Datatypes

- ✓ Strings
- ✓ Lists
- ✓ Sets
- ✓ Sorted Sets
- ✓ Hashes
- ✓ Bitmaps
- ✓ Hyperlogs
- ✓ Geospatial Indexes

PowerBI





Why PowerBI:

- Free PowerBI Desktop development tool
- Economical licensing model (only \$10 per user per month)
- Accessible on all platforms - Android, Mac, Iphone, Windows (However, development can only be done on windows)
- Easy to learn
- Cloud tool (Software as a service). No Infrastructure required
- Ability to write python and R scripts for data transformation
- Can connect to practically any datasource



How does it work:

- Developer creates the dashboard using PowerBI desktop tool on a windows machine (Developer tool is free and does not require any license)
- 'Developer publishes the dashboard on PowerBI service Cloud (This requires a powerbi license @ \$10/user/month)
- End users can access the dashboards from any machine using a web browser. (All end users also need a powerbi license @ \$10/user/month)
- Dashboard contains the dataset embedded in it
-

File Home Insert Modeling View Help Format Data / Drill

Paste
 Cut
 Copy
 Format painter

Get data

Excel
 Power BI datasets
 SQL Server
 Enter data
 Recent sources

Transform data
 Refresh

New visual
 Text box
 More visuals

New measure
 Quick measure

Publish

ClipboardDataQueriesInsertCalculationsShare

Total Row Count

46.65K

Total Games on Clay

2361

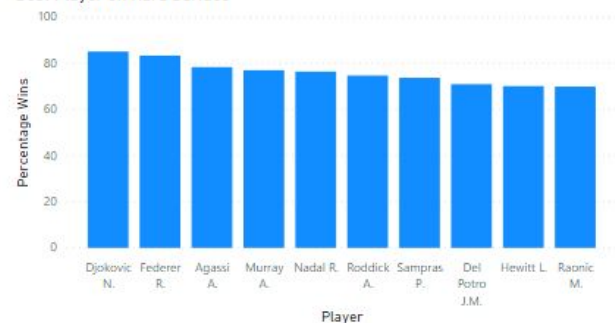
Total Games on Hard Surface

4155

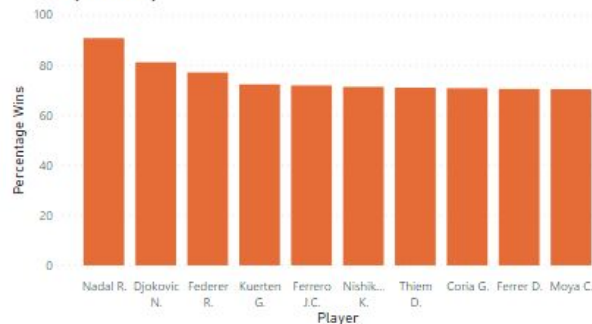
Total Games on Grass

925

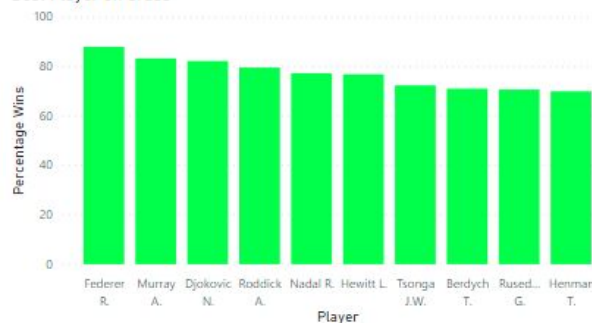
Best Player on Hard Surface



Best Player on Clay



Best Player on Grass



Filters

Visualizations

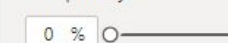


Search

Color



Transparency



Revert to default

Lock aspect... Off

General

Fields

Search

Data

top_clay

top_grass

top_hard

☐ Column1

☐ Count_Los

☐ Count_Win

☐ perc_win

☐ Player

☐ Surface

☐ total_play

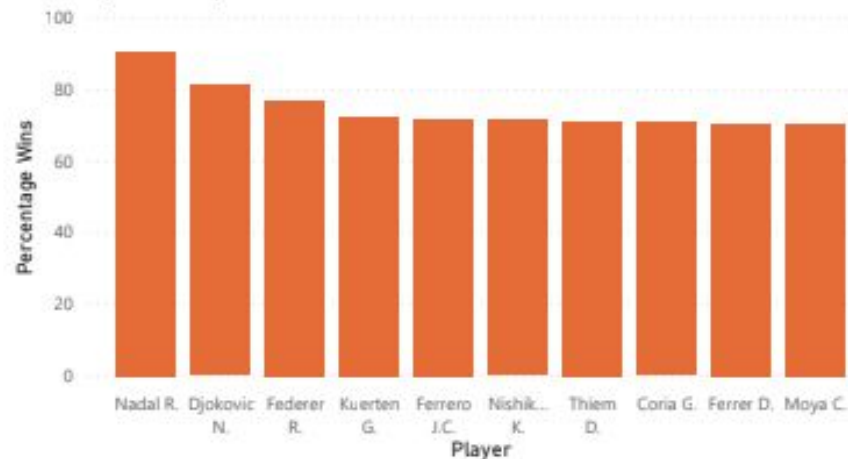
Total Row Count

46.65K

Total Games on Clay

2361

Best Player on Clay



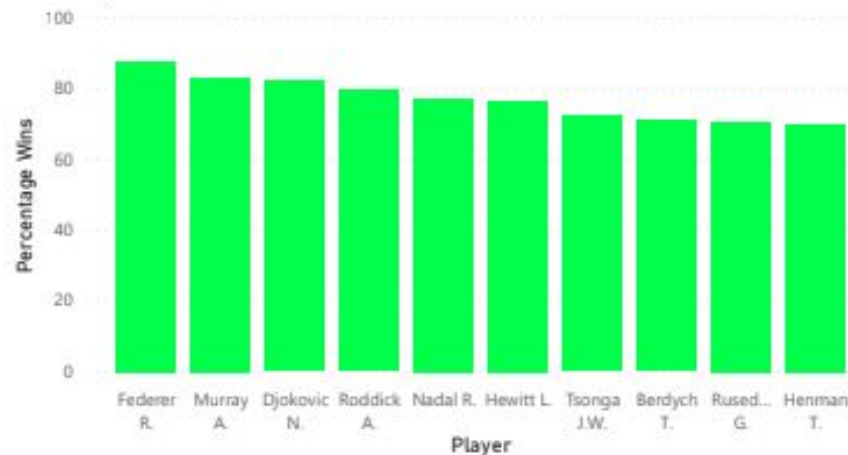
Total Games on Hard Surface

4155

Total Games on Grass

925

Best Player on Grass



Best Player on Hard Surface

