Apache Spark and Scala

Module 1: Getting Started / Introduction to Scala

Course Topics

Module 1

Getting Started / Introduction to Scala

Module 2

Spark and Big Data

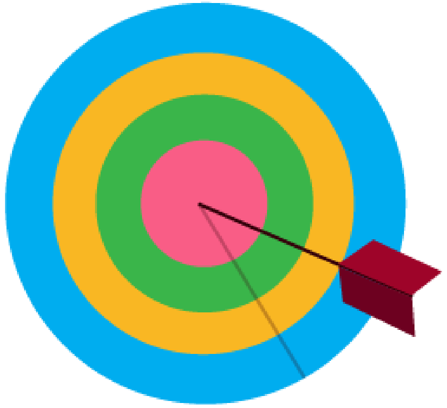
Module 3

Understanding RDDs

Session Objectives

This session will help you to understand:

- Big Data
- IBM's Big Data Definition
- Some Big Data Examples
- Sparks Basics
- Why Spark ?
- Spark Components
- Scala Basics
- Why Scala ?
- Users of Scala
- Scala Frameworks
- Scala Usage
- Software Installation
- Scala Hands-on
- Scala community



Introduction to Big Data

THREAT Toons™
by: Alex Savchuk
ALL OF THIS BIG DATA... AND WHAT AM I SUPPOSED TO DO WITH IT?
DATA

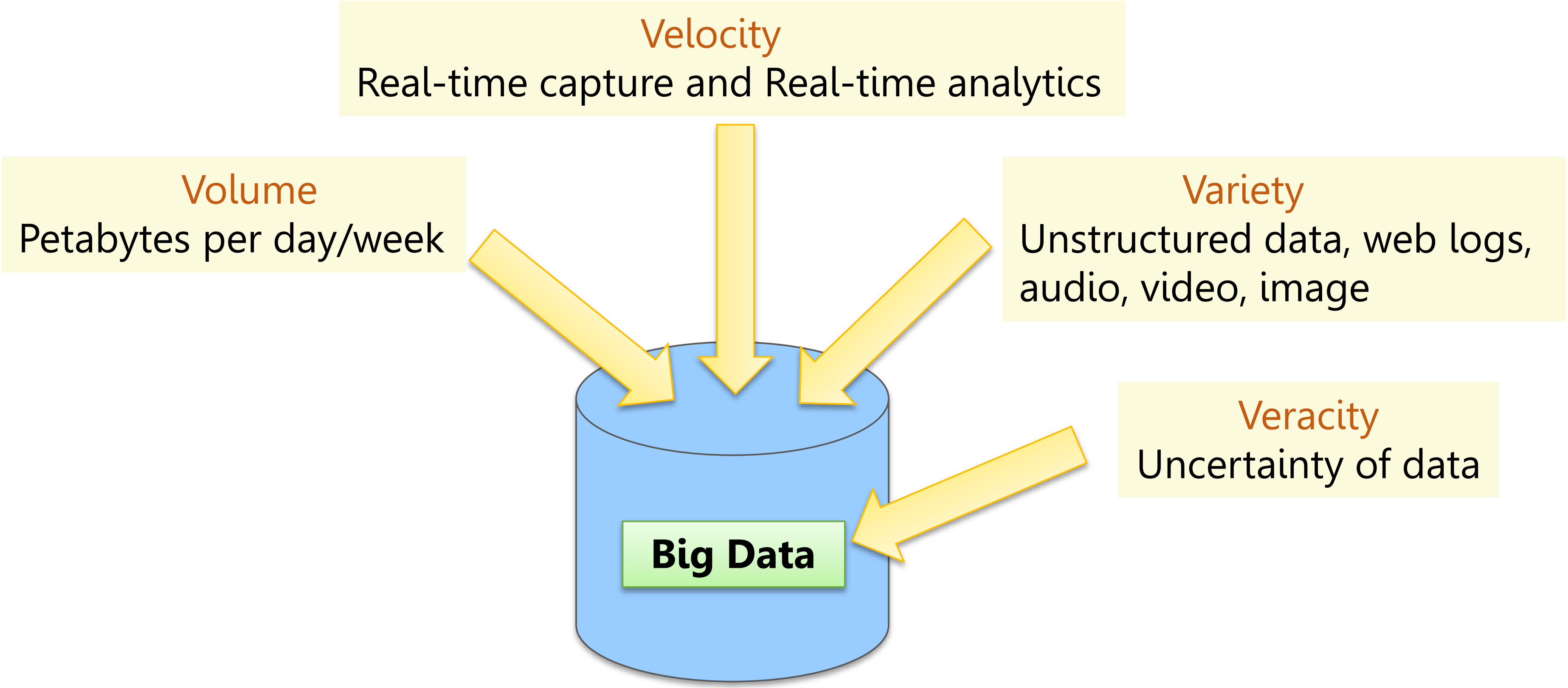
Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate

The challenges of big data includes: analysis, capture, data curation, search, sharing, storage, transfer, visualization, and information privacy

IBM's Big Data Definition

skillspeed
for the serious learner

- IBM's Definition – Big Data Characteristics
- http://www.ibmbigdatahub.com/infographic/four-vs-big-data/



Velocity
Real-time capture and Real-time analytics

Volume
Petabytes per day/week

Variety
Unstructured data, web logs, audio, video, image

Veracity
Uncertainty of data

Some Big Data Examples

skillspeed
for the serious learner

- NYSE broadcasts several levels of data, including trade prices, sizes
- NYSE Technologies receives four to five terabytes of a data in a day and which is used for complex analytics, market surveillance, capacity planning and monitoring

NYSE generates about one terabyte of new trade data per day to perform stock trading analytics to determine trends for optimal trades

Check your Understanding – 1

Which of the following are the Big Data Solutions Candidates?

a) Processing 1.5 TB data everyday

b) Processing 30 minutes Flight sensor data

c) Interconnecting 50K data points (approx. 1 MB input file)

d) Processing User clicks on a website

Check your Understanding – Solution

Which of the following are the Big Data Solutions Candidates?

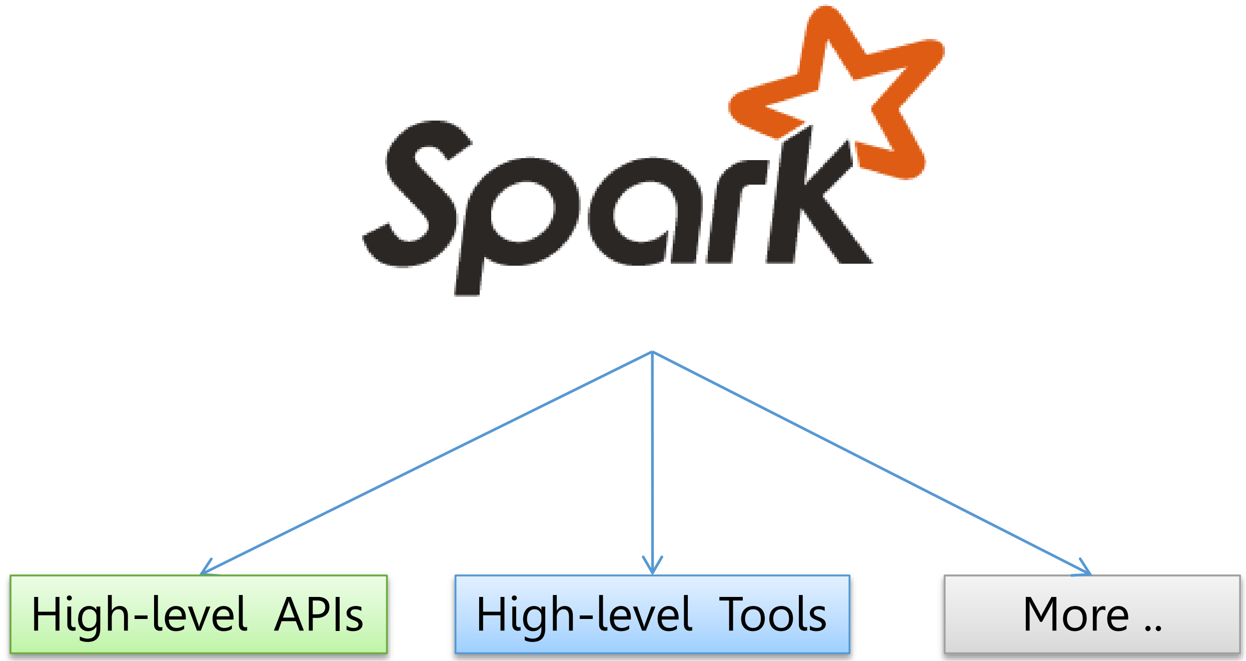
- ✓ a) Processing 1.5 TB data everyday
- ✓ b) Processing 30 minutes Flight sensor data
- ✓ c) Interconnecting 50K data points (approx. 1 MB input file)
- ✓ d) Processing User clicks on a website

ALL of the options are Big Data solutions Scenario. Even if the input size of the problem is small, the processing might make the scenario as Big Data Problem

Introduction to Spark

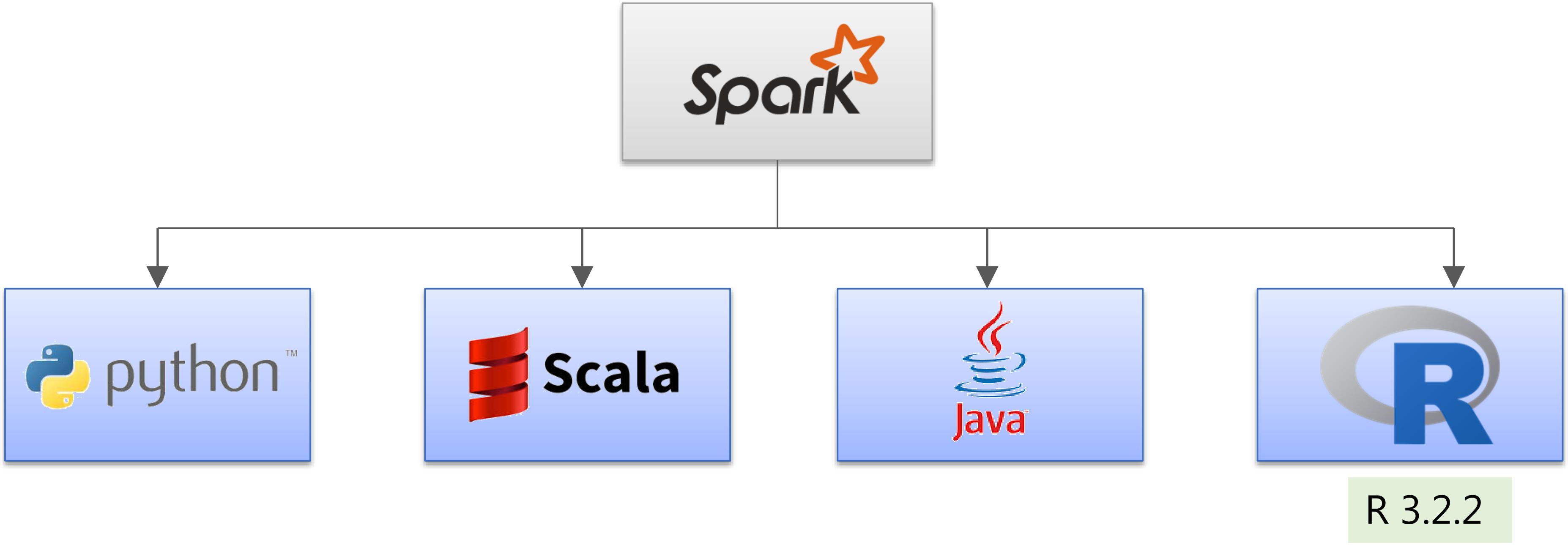
Spark Basics

- Apache Spark is a general-purpose cluster in-memory computing system which is used for data analytics
- It provides high-level APIs in Java, Scala and Python and an optimized engine that supports general execution graphs
- Apache Spark Provides various high level tools like Spark SQL for structured data processing, R programming Language for analyzing large datasets and MLlib for Machine Learning etc.

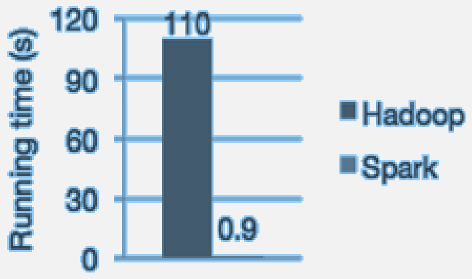


Spark Basics (Cont’d)

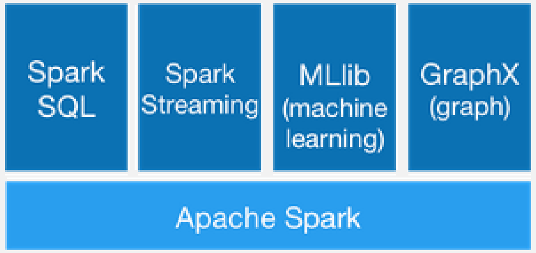
Spark framework is polyglot – It can be programmed in several programming languages (Java, Scala ,R 3.2.2 and Python supported)



Why Spark?



Speed
Run programs up to 100x faster than Hadoop Map Reduce in memory



Spark SQL

Spark Streaming

MLlib (machine learning)

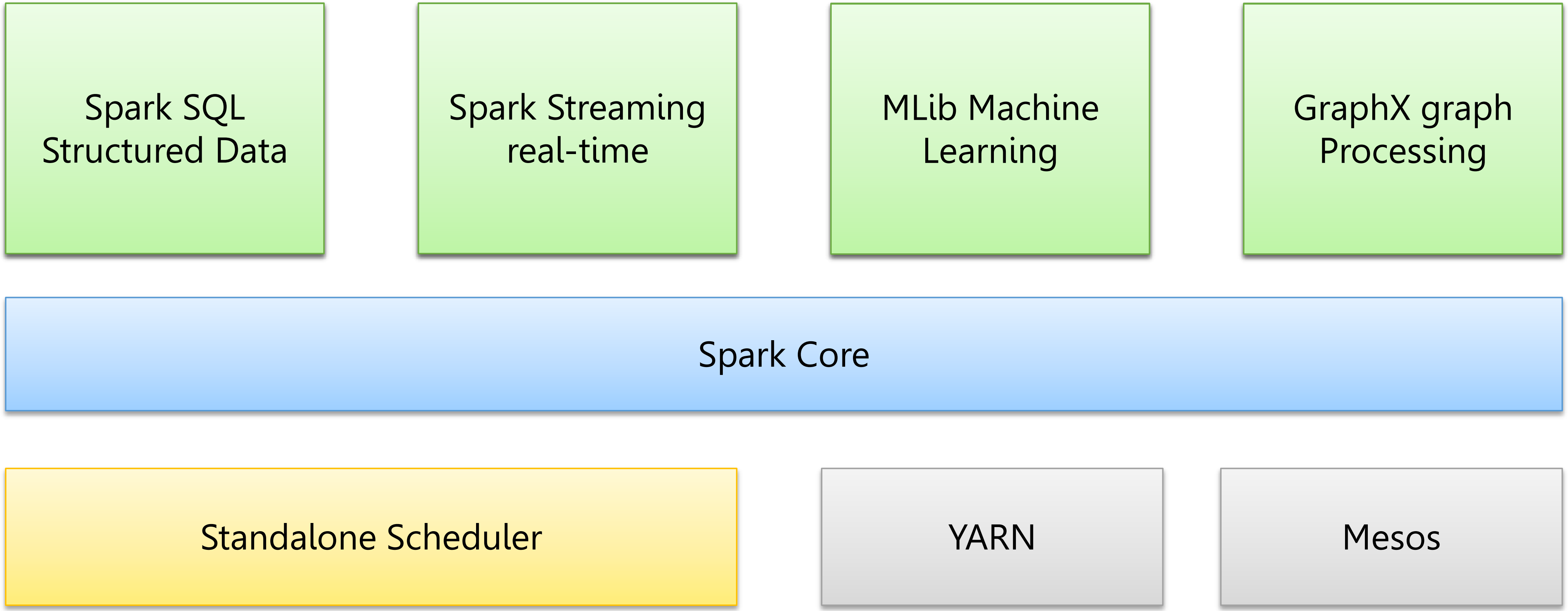
GraphX (graph)

Apache Spark

Generality
Combine SQL ,streaming and complex analytics into one platform

Runs Everywhere
Spark runs on Hadoop, Mesos.standalone or in cloud

Spark Components



Spark SQL Structured Data

Spark Streaming real-time

MLib Machine Learning

GraphX graph Processing

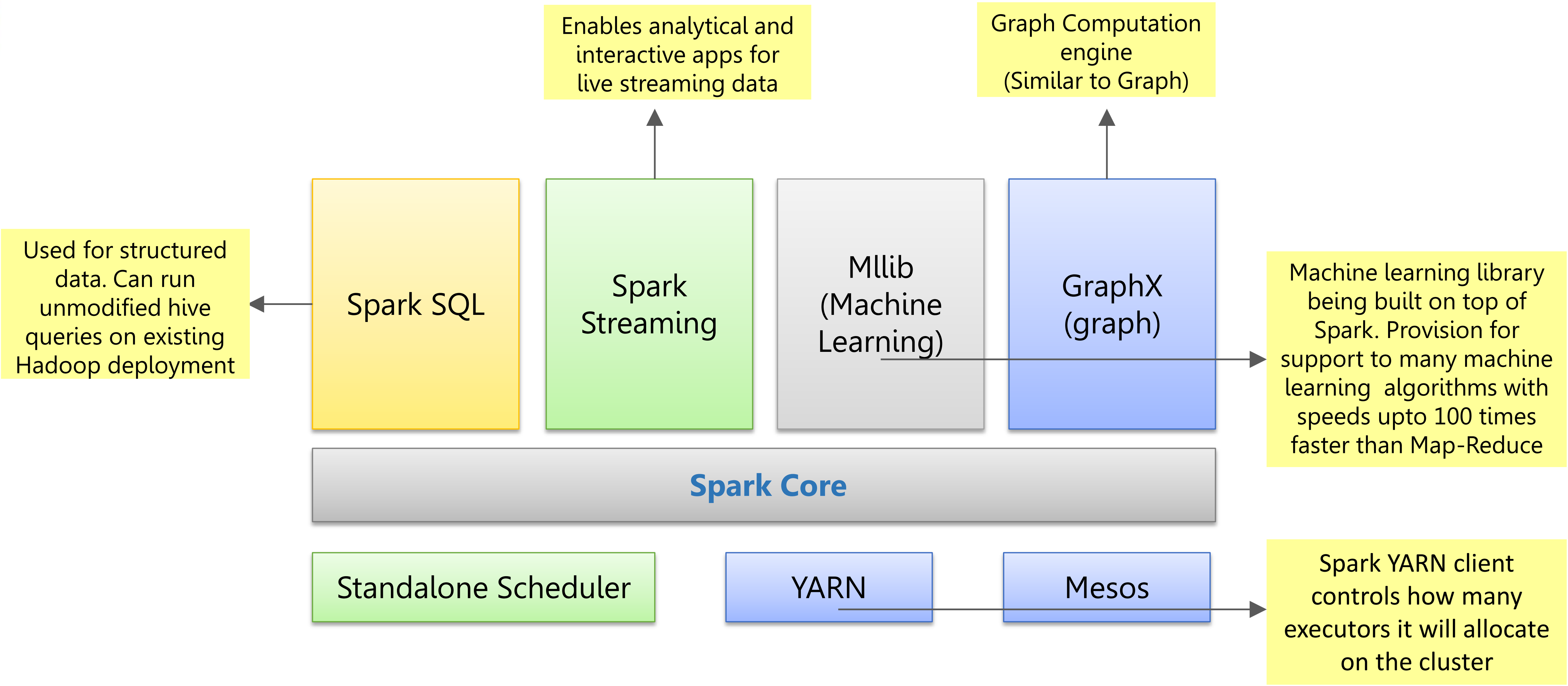
Spark Core

Standalone Scheduler

YARN

Mesos

Spark Components (Cont'd)



Enables analytical and interactive apps for live streaming data

Graph Computation engine (Similar to Graph)

Used for structured data. Can run unmodified hive queries on existing Hadoop deployment

Spark SQL

Spark Streaming

Mllib (Machine Learning)

GraphX (graph)

Machine learning library being built on top of Spark. Provision for support to many machine learning algorithms with speeds upto 100 times faster than Map-Reduce

Spark Core

Standalone Scheduler

YARN

Mesos

Spark YARN client controls how many executors it will allocate on the cluster

Spark + Scala = CREATES MAGIC

Introduction to Scala

Martin Odersky and his team started developing Scala in 2001

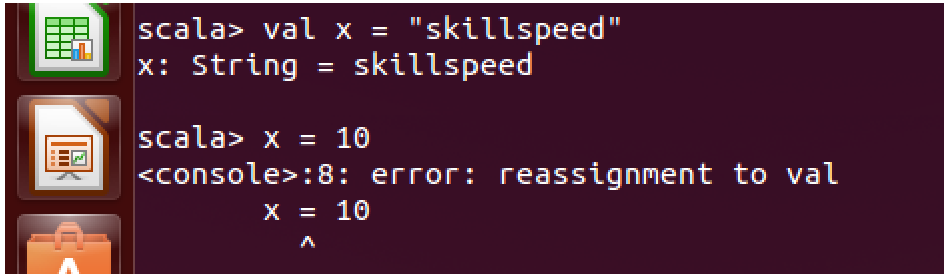
Introduction to Scala (Cont'd)

- Scala is a general purpose programming language, multiparadigm object oriented, functional, scalable
- Aimed to implement common programming patterns in a concise, elegant, and type-safe way
- Supports both object-oriented and functional programming styles, thus helping programmers to be more productive
- Publicly released in January 2004 on the JVM platform and a few months later on the .NET platform

Introduction to Scala (Cont’d)

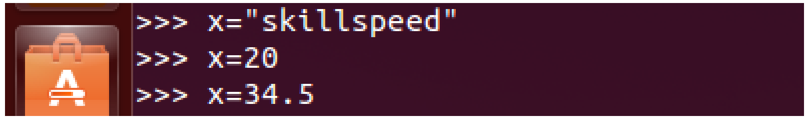
Scala is Statically Typed

- Statically typed language binds the type to a variable for its entire scope



```
scala> val x = "skillspeed"
x: String = skillspeed

scala> x = 10
<console>:8: error: reassignment to val
       x = 10
         ^
```

- Dynamically typed languages bind the type to the actual value referenced by a variable .Example : python



```
>>> x="skillspeed"
>>> x=20
>>> x=34.5
```

- Fully supports Object Oriented Programming
- Everything is an object in Scala
- Unlike Java, Scala does not have primitives
- Supports “static” class members through Singleton Object Concept
- Improved support for OOP through Traits

Why Scala?

- Scala is pure object-oriented language. Conceptually, every value is an object and every operation is a method-call
- Scala is also a functional language and supports immutable data structures
- Many big data technologies use Scala like Spark, Kakfka, Storm, Akka, Scalding and web frameworks like Play

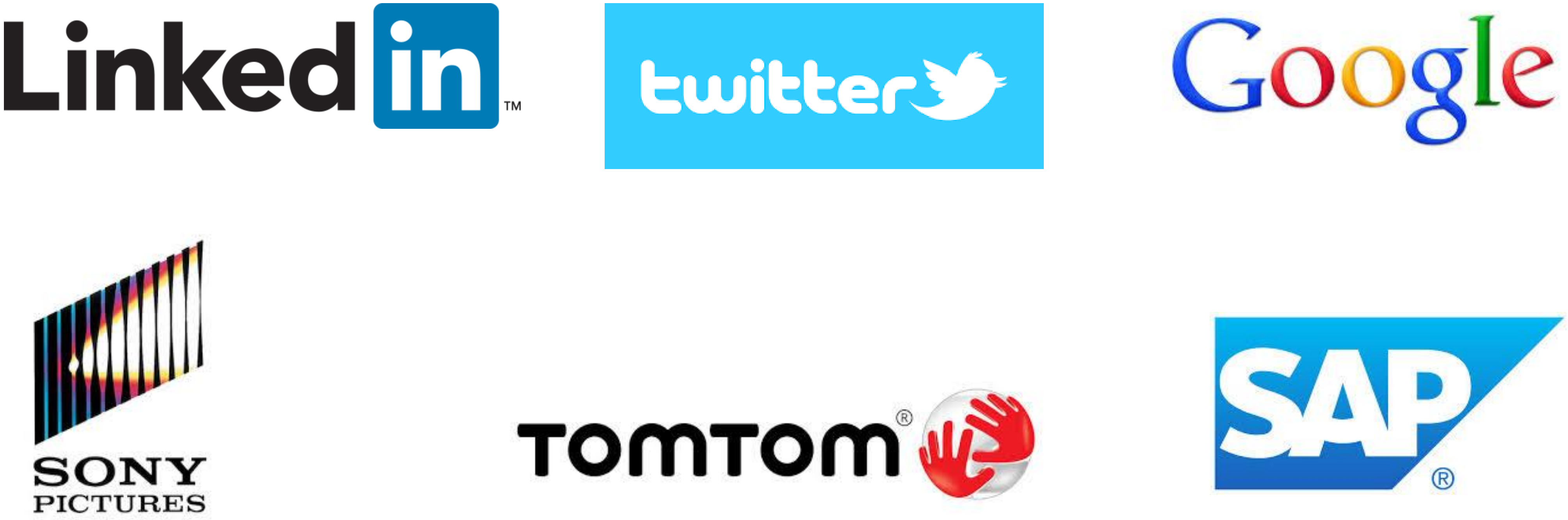
Why Scala? (Cont'd)

Scala code compared to Java code

Java	Scala

Java Code
List<String> list = new ArrayList<String>();
list.add("1");
list.add("2");
list.add("3");
	Scala Code
	val list = List("1", "2", "3")

Scala Users



Scala Frameworks

Play – For Web Development

Play is a high-productivity Java and Scala web application framework that integrates the components and APIs you need for modern web application development

Scalding – For Map/Reduce

Scalding is a Scala library that makes it easy to specify Hadoop MapReduce jobs. Scalding is built on top of Cascading, a Java library that abstracts away low-level Hadoop details

Akka – Actors Based Framework

Akka is a toolkit and runtime for building highly concurrent, distributed, and fault tolerant applications on the JVM. Akka is written in Scala

Scala Frameworks (Cont'd)

Spark – In – memory Processing

Apache Spark is a general-purpose cluster in-memory computing system. It is used for fast data analytics and it abstracts APIs in Java, Scala and Python, and provides an optimized engine that supports general execution graphs

Apache Kafka

Apache Kafka is publish-subscribe messaging rethought as a distributed commit log

Scala Usage

Scripting

Web Application

Messaging

Mobile Android Apps

Digital Subscriber Line

GUI (Graphical User Interface)

Check your Understanding – 2

Which Features are supported by Scala?

a) Less error prone functional style

b) High maintainability and productivity

c) High scalability

d) High testability

e) Provides features of concurrent programming

Check your Understanding – Solution

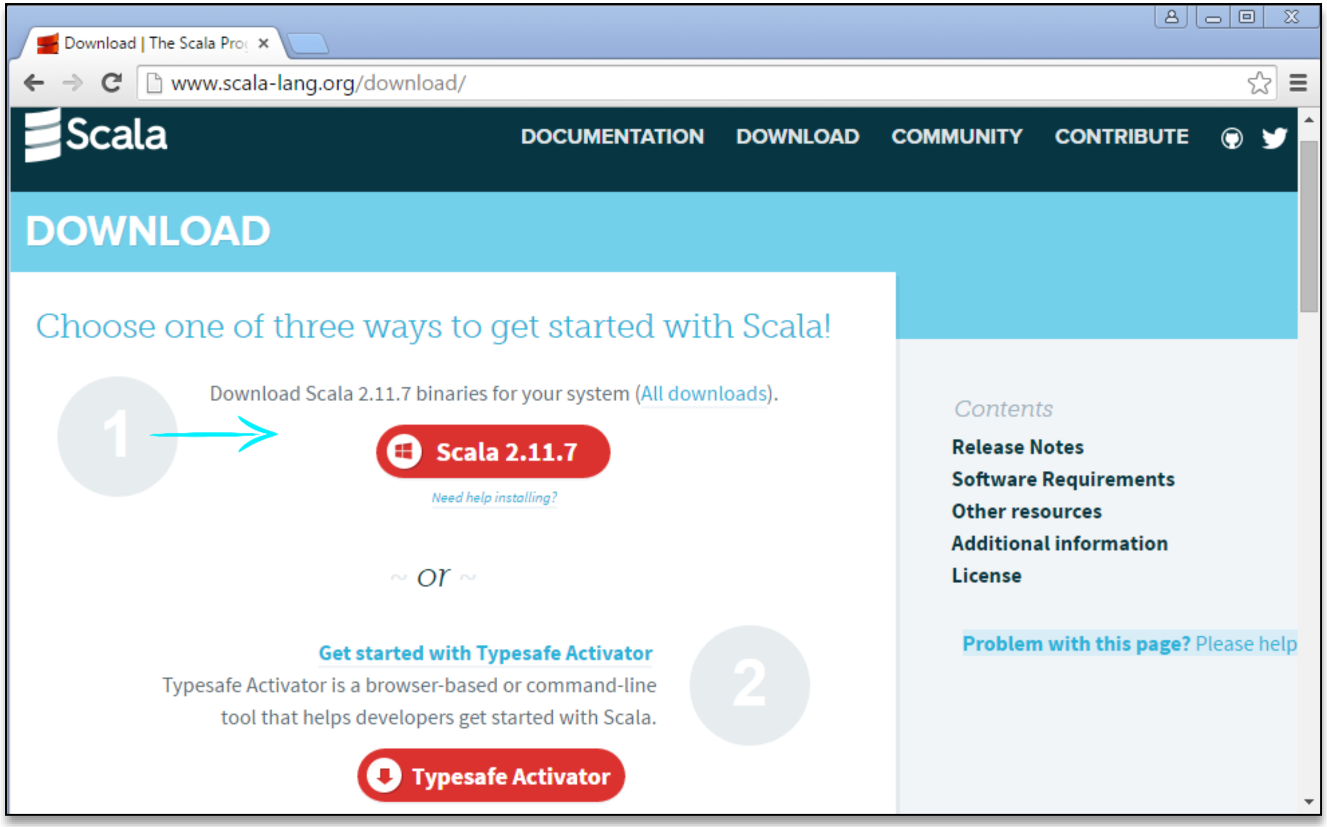
Which Features are supported by Scala?

- ✓ a) Less error prone functional style
- ✓ b) High maintainability and productivity
- ✓ c) High scalability
- ✓ d) High testability
- ✓ e) Provides features of concurrent programming

All of these are the features of Scala

Software Installation

- Latest version can be downloaded from: http://www.scala-lang.org/download/
- Install the Scala and Set the Scala Path in Machine



DOWNLOAD

Choose one of three ways to get started with Scala!

Download Scala 2.11.7 binaries for your system (All downloads).

Get started with Typesafe Activator

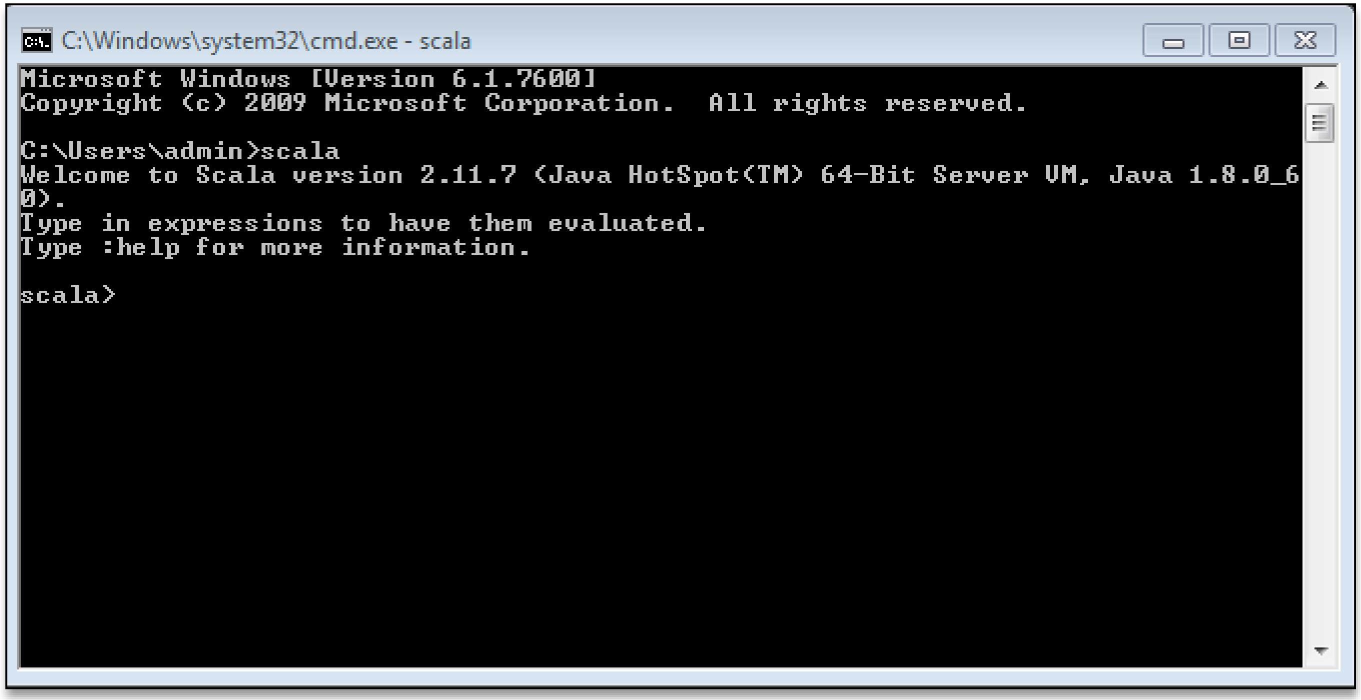
Typesafe Activator is a browser-based or command-line tool that helps developers get started with Scala.

Problem with this page? Please help

Note: Extensive installation Guide is available in LMS

Scala Hands-on (Cont'd)

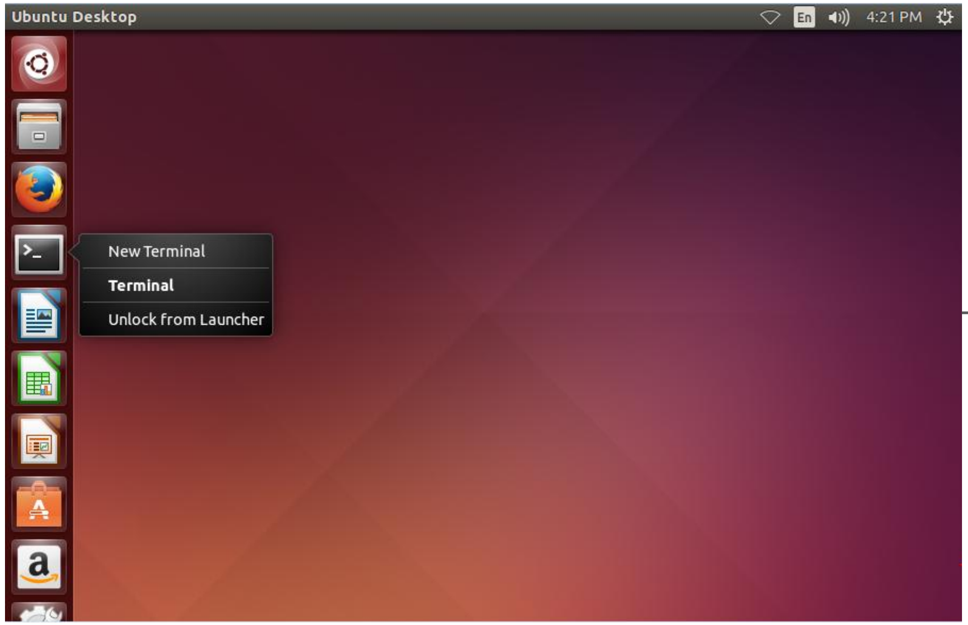
- Start → Type Run → Type cmd
- Type scala

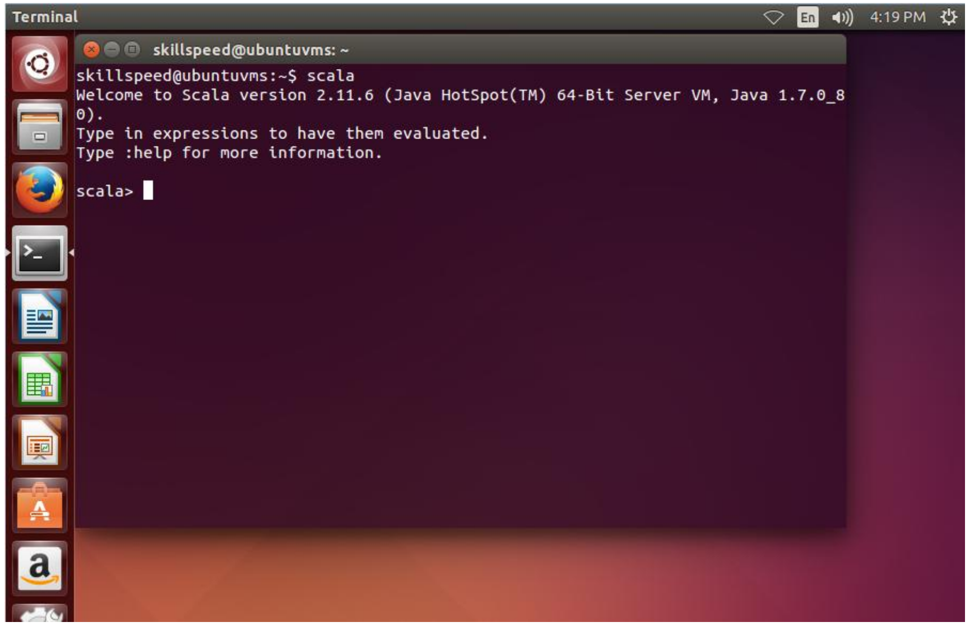


```
Microsoft Windows [Version 6.1.7600]
Copyright (c) 2009 Microsoft Corporation.  All rights reserved.

C:\Users\admin>scala
Welcome to Scala version 2.11.7 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_6
0).
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

Scala Hands-on (Cont'd)

- Download Vmware Player
- Double-click on Vmware Player Workstation → Open the Virtual Machine (it will open the Ubuntu desktop)
- Install scala
- Then select New Terminal → It will open the Terminal and install Scala then type scala



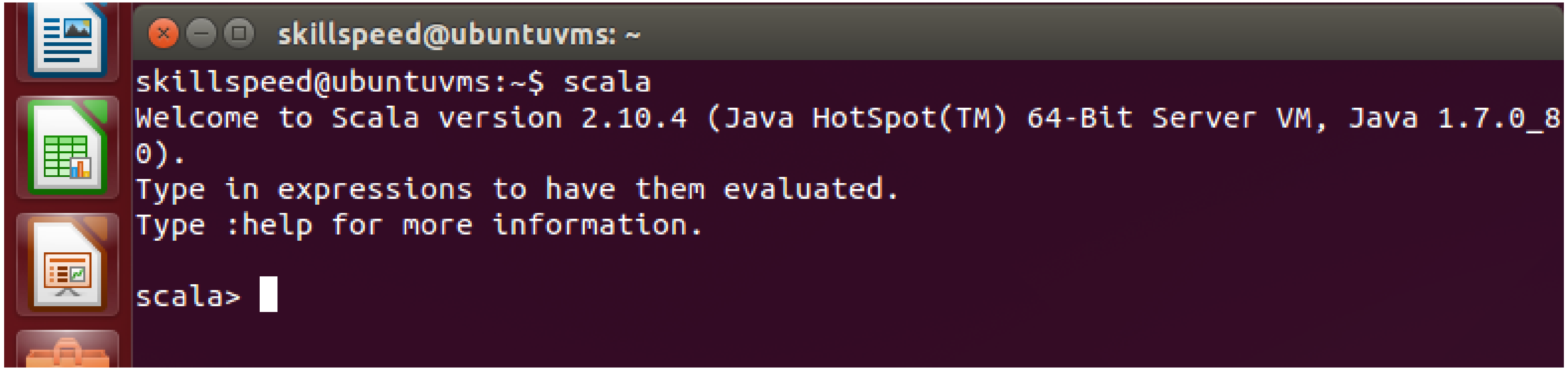


```
skillspeed@ubuntuvms:~$ scala
Welcome to Scala version 2.11.6 (Java HotSpot(TM) 64-Bit Server VM, Java 1.7.0_8
0).
Type in expressions to have them evaluated.
Type :help for more information.

scala> 
```

Note: Installation Guide for Linux is Available in LMS

Scala Hands-on (Cont'd)

- REPL: Read - Evaluate - Print - Loop
- Easiest way to get started with Scala, acts as an interactive shell interpreter
- Even though it appears as interpreter, all typed code is converted to Bytecode and executed
- Invoked by typing Scala as shown below



```
skillspeed@ubuntuvms:~$ scala
Welcome to Scala version 2.10.4 (Java HotSpot(TM) 64-Bit Server VM, Java 1.7.0_8
0).
Type in expressions to have them evaluated.
Type :help for more information.

scala> 
```

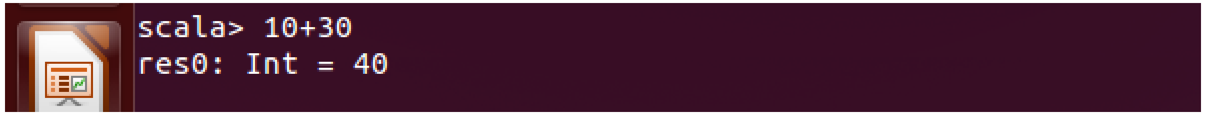
Scala Hands-on (Cont'd)

After you type an expression, such as 10 + 30, and hit enter:

```
scala> 10 + 30
```

The interpreter will print:

```
res0: Int = 40
```



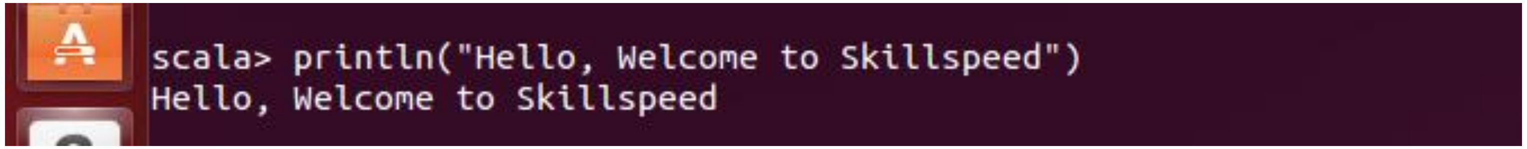
```
scala> 10+30
res0: Int = 40
```

This line includes:

- An automatically generated or user-defined name to refer to the computed value (res0, which means result 0),
- A colon (:), followed by the type of the expression (Int),
- An equals sign (=),
- The value resulting from evaluating the expression (30)

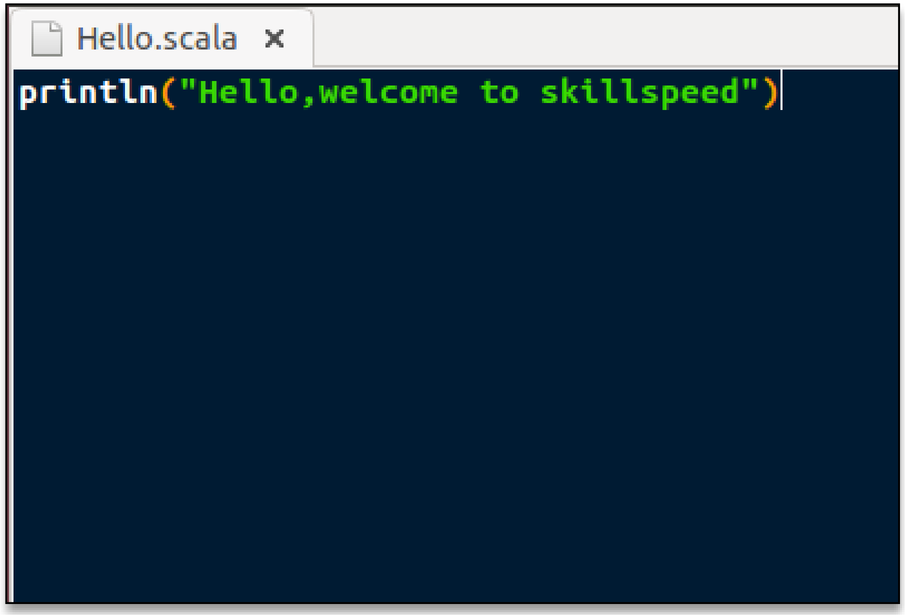
Scala Hands-on (Cont'd)

In the beginning, you started with the REPL



```
scala> println("Hello, Welcome to Skillspeed")
Hello, Welcome to Skillspeed
```

- Scala scripts can be written in text files and saves the script with a .scala extension
- It indicates to the operating system and programmer that the file is actually a scala program

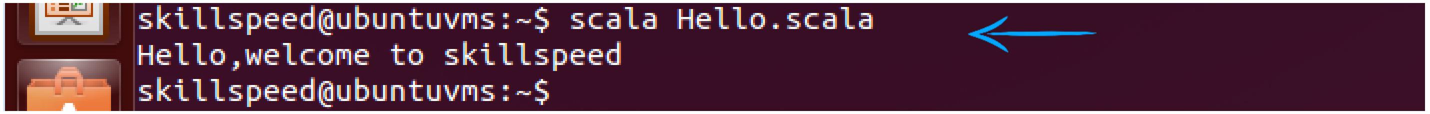


```
println("Hello,welcome to skillspeed")
```

Scala Hands-on (Cont’d)

The scripts can be read into the interpreter in several ways:

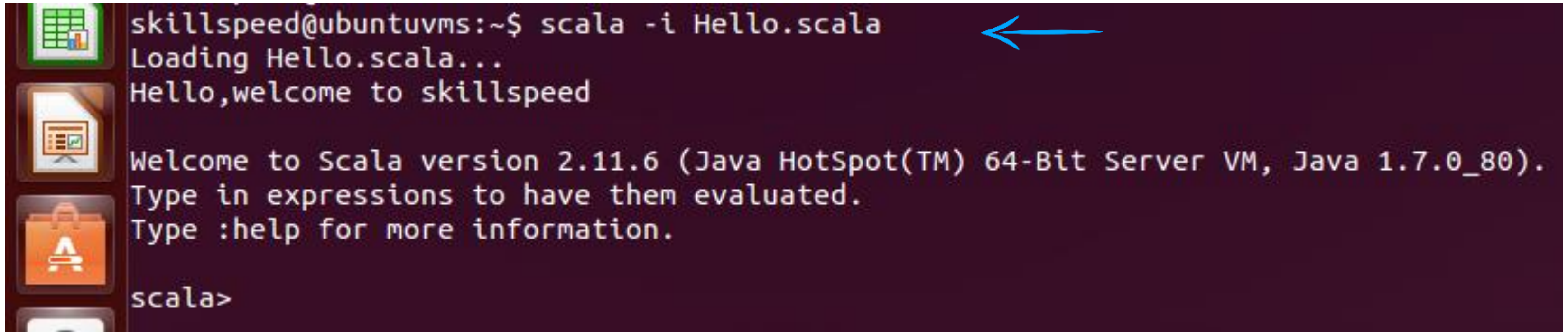
scala Hello.scala   # here Hello is Script name



```
skillspeed@ubuntuvms:~$ scala Hello.scala
Hello,welcome to skillspeed
skillspeed@ubuntuvms:~$
```

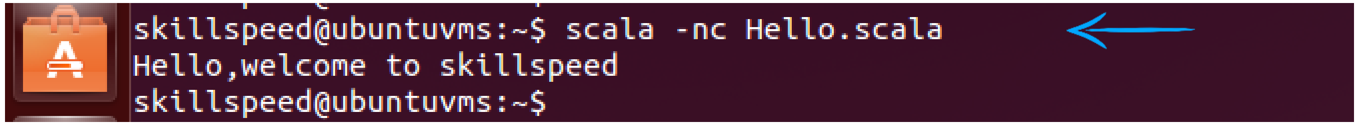
The script is executed and the REPL is immediately closed

- scala –i Hello.scala   (Output prints and opens the scala REPL)



```
skillspeed@ubuntuvms:~$ scala -i Hello.scala
Loading Hello.scala...
Hello,welcome to skillspeed

Welcome to Scala version 2.11.6 (Java HotSpot(TM) 64-Bit Server VM, Java 1.7.0_80).
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

- scala –nc Hello.scala



```
skillspeed@ubuntuvms:~$ scala -nc Hello.scala
Hello,welcome to skillspeed
skillspeed@ubuntuvms:~$
```

Scala Community

Developers in countries all over the world are using Scala for a large variety of applications across a broad range of industries

Popular ways to connect with the Scala community are via mailing lists or IRC channels

Though there are plenty of opportunities to connect face-to-face with others in the community– for example, via local Scala Meetups, or local Scala user groups

Check your Understanding – 3

Scala REPL acts as scala Interpreter

a) True

b) False

Check your Understanding – Solution

Scala REPL acts as scala Interpreter

a) True

✓ b) False

False

Check your Understanding – 4

Scala supports primitive and wrapper classes ?

a) True

b) False

Check your Understanding – Solution

Scala supports primitive and wrapper classes ?

a) True

[✓ b) False]

False

Questions

thank you!