Scala, Big Data, dan Apache Spark

Bambang Purnomosidi D. P.
Web Intelligence Research Group
STMIK AKAKOM
http://github.com/wi-rg

Scala: Scalable Language (and more)

- Dibuat oleh Martin Odersky - École Polytechnique Fédérale de Lausanne.
- Paradigam: functional, object-oriented, imperative, concurrent.
- Deployment target: JVM (main target), JavaScript (Scala.js - stable), LLVM (Scala Native – experiment)
- Type system: static, strong, inferred, structural
- Java interoperability: Scala → Java, Java → Scala
- Jangan bertanya “Apakah Scala dipakai di industri?” - kecuali tidak bisa memakai Google.
- Actor model – let it crash, dulu masuk pada standard library – sekarang dikembangkan di Akka (http://www.akka.io).

Functional Programming di Scala

- Functional programming: komputasi merupakan evaluasi dari fungsi (pada matematika) serta menghindari *changing-state* dan *mutable data*.
  - Everything is an expression
  - Mutable and immutable data
  - Type inference
  - Anonymous function / closure
  - Lazy evaluation (fungsi tidak akan dievaluasi jika belum diperlukan)
  - Higher-order functions (fungsi yang bisa menjadi parameter fungsi)
  - Currying (fungsi yang mengambil lebih dari satu parameter menjadi serangkaian fungsi – masing2 mempunyai satu parameter).
  - Pattern matching

Scala Ecosystem

- IDE: Scala IDE, vim plugins, Emacs packages, Netbeans Scala, IntelliJ IDEA.
- Pustaka: Maven repos (private atau public), The Scala Library Index (http://index.scala-lang.org).
- Build tools: sbt (officially maintained), Gradle, Maven.
- Concurrent and distributed programming: Akka
- Testing: specs2, ScalaTest, ScalaCheck
- Dependency Injection: MacWire, Scaldi
- Web: Play framework, Scalatra
- I/O and Microservices: Colossus (tumblr)
- RPC: finagle (twitter).

Scala dan Data Access

- SQL: Slick (http://slick.lightbend.com). Database supported: DB2 (via slick-extensions), Derby/JavaDB, H2, HSQLDB/HyperSQL, Microsoft SQL Server (via slick-extensions), MySQL, Oracle (via slick-extensions), PostgreSQL, SQLite.
- NoSQL: Kebanyakan mendukung Java dan / atau Scala. When no Scala specific driver available: use Java driver.

Scala dan Big Data

- Big data: data set yang sedemikian besar dan kompleks.
- Big data secara minimal melibatkan kemungkinan pemerolehan data (secara tradisional: data warehousing), penyimpanan, analisis, pencarian pattern.
- Data: persistent dan stream.
- Persistent: Hadoop + Cascading. Scala => Scalding.
- Stream: Complex Event Processing dan Stream Processing

Apache Spark

- Apache Spark is a fast and general-purpose cluster computing system. It provides high-level APIs in Java, Scala, Python and R, and an optimized engine that supports general execution graphs. It also supports a rich set of higher-level tools including Spark SQL for SQL and structured data processing, MLlib for machine learning, GraphX for graph processing, and Spark Streaming

Apache Spark overview

(http://spark.apache.org/docs/latest/)

Komponen Apache Spark

- Cluster: stand alone, Apache Mesos, Hadoop YARN
- Applications: dibangun menggunakan Scala, Java, Python, R. Aplikasi dibangun dan di-submit ke cluster untuk pemrosesan.
- Data set (RDD – Resilient Distributed Dataset)
- Modules. Default yang ada pada Apache Spark:
  - Spark SQL: untuk structured data
  - Spark Streaming: untuk streaming
  - Spark Mllib: pustaka untuk ML
  - Spark GraphX: pustaka untuk pengolahan graph dan paralel graph.