Big Data: Data Wrangling Boot Camp
BD Tools and Techniques

Chuck Cartledge, PhD

16 September 2016
Table of contents I

1 Introduction
2 Amdahl
3 BD Processing
4 Languages
5 Q & A
6 Conclusion
7 References
8 Concepts
What are we going to cover?

We’re going to talk about:

- Why it is important to understand your problem
- What are single and multithreaded programs
- What are different tools, and frameworks to support BD processing
- What languages and programming paradigms fit the BD world
- A passing appreciation of BD and Chaos concepts
A little math

Amdahl’s Law [2]

- Time for serial execution \( T(1) \)
- Portion that can NOT be paralyzed \( B \in [0, 1] \)
- Number of parallel resources \( n \)

\[
T(n) = T(1) \times (B + \frac{1}{n}(1-B))
\]

\[
S(n) = \frac{T(1)}{T(n)} = \frac{1}{B + \frac{1}{n}(1-B)}
\]

Dr. Gene Amdahl (circa 1960)
Amdahl’s Law (A summary)

Division and measurement of serial and parallel operations appears time and again. (Shades of Mandelbrot.)

- “Make the common fast.”
- “Make the fast common.”
- Understand what parts have to be done serially
- Understand what parts can be done in parallel

Need to factor in “overhead” costs when computing speed up.
Some questions are easily stated, . . .

Which of these are paralizable (and why)?

1. \(a[i] = b[i] + c[i]\)
2. \(a[i] = f(b)\)
3. \(a[i] = a[i - 1] + b[i - 1]\)
4. \(a = b + c\)
Single thread vs. multithreads

- Single-threaded process – has full access to CPU and RAM
- Multithreaded process – shares access to CPU and RAM
- Multithreaded makes sense with independent tasks
- Multithreaded may share the same memory space (language dependent)

Coordination across multiple threads can be tricky.

Image from [3].
Hadoop multithreading hidden from view.

Image from [5].
An overview

Vocabulary

- Data Sources – where data comes from
- Ingestion – how data is pre-processed for acceptance
- Data Sea/Lake – where data lives
- Processing – how data is processed prior to storage
- Data warehouse – transition from SQL to NoSQL
- Analysis – extracting information from data
- User interface – how the user interacts with the information
An overview

Big Data Landscape Q3/2016

Polyglott Persistence & Analytics
NewSQL: SQL meets NoSQL.
In-Memory Databases: Fast access.
Index Machines: Fast aggregation and search.

Structured Data Lake: The eternal memory.
Efficient data serialization formats:
- Integrated compression
- Column-oriented storage
- Predicate pushdown

Micro Analytics Services: Substitute for reporting servers.
Dashboards:
Charting Libraries:

User Interface
Data Lake
Data Warehouse
Ingestion
Processing
Data Science Interactive Analysis Reporting & Dashboards

Data Sources
Analytics

Data Workflows ETL Jobs Massive Parallelization Data Logistics Stream Processing

Image from [1].
Another collection of Open Source BD tools

Tools partitioned differently:

- Big Data search
- Business Intelligence
- Data aggregation
- Data Analysis & Platforms
- Databases / Data warehousing
- Data aggregation
- Data Mining
- Document Store
- Graph databases
- Grid Solutions
- In-Memory Computing
- KeyValue
- Multidimensional
- Multimodel
- Multivalue database
- Object databases
- Operational
- Programming
- Social
- XML Databases

Image from [10].
The reference has links to each piece of software.
Each is different for a reason

Hammer and nails . . .

“...it is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail.”

Abraham H. Maslow [8]

Image from [11].
Each is different for a reason

A simple comparison of some languages

Languages are created by people to solve certain types of problems.

- C# – declarative, functional, generic, imperative, object-oriented (class-based)
- Java – client side, compiled, concurrent, curly-bracket, impure, imperative, object-oriented (class-based), procedural, reflective
- Python – compiled, extension, functional, imperative, impure, interactive mode, interpreted, iterative, metaprogramming, object-oriented (class-based), reflective, scripting
- R – array, impure, interpreted, interactive mode, list-based, object-oriented prototype-based, scripting

Categorizations from [12].
Each is different for a reason

Vocabulary (1 of 2)[12].

- **array** – generalize operations on scalars to apply transparently to vectors, matrices, and higher-dimensional arrays.
- **client side** – languages are limited by the abilities of the browser or intended client.
- **compiled** – languages typically processed by compilers, though theoretically any language can be compiled or interpreted.
- **concurrent** – languages provide language constructs for concurrency.
- **curly-bracket** – languages have a syntax that defines statement blocks using the curly bracket or brace characters.
- **declarative** – languages describe a problem rather than defining a solution.
- **extension** – languages embedded into another program and used to harness its features in extension scripts.
- **functional** – languages define programs and subroutines as mathematical functions.
- **generic** – language is applicable to many domains.
- **imperative** – languages may be multi-paradigm and appear in other classifications.
- **impure** – languages containing imperative features.
- **interactive mode** – languages act as a kind of shell.
Vocabulary (2 of 2)[12].

- interpreted – languages are programming languages in which programs may be executed from source code form, by an interpreter.
- iterative – languages are built around or offering generators.
- list-based – languages are a type of data-structured language that are based upon the list data structure.
- metaprogramming – hat write or manipulate other programs (or themselves) as their data or that do part of the work that is otherwise done at run time during compile time.
- object-oriented (class-based) – support objects defined by their class.
- object-oriented prototype-based – languages are object-oriented languages where the distinction between classes and instances has been removed
- procedural – languages are based on the concept of the unit and scope
- reflective – languages let programs examine and possibly modify their high level structure at runtime.
- scripting – another term for interpreted
Each reflects/supports a programming paradigm

A plethora of programming paradigms:

- Action
- Dynamic
- Event-driven
- Generic
- Imperative
- Language-oriented
- Parallel
- Declarative
- Semantic
- Functional
- Structured

Image from [13].

Each paradigm is a result of a problem domain.
Each is different for a reason.

Same image.

Image from [13].
What does the future hold?

“If languages are not defined by taxonomies, how are they constructed? They are aggregations of features. Rather than study extant languages as a whole, which conflates the essential with the accidental, it is more instructive to decompose them into constituent features, which in turn can be studied individually. The student then has a toolkit of features that they can re-compose per their needs.”

S. Krishnamurthi [6]

New languages will be created all the time to fit needs.
Q & A time.

“The Answer to the Great Question . . . Of Life, the Universe and Everything . . . is . . . forty-two,’ said Deep Thought, with infinite majesty and calm.”

Douglas Adams, The Hitchhiker’s Guide to the Galaxy
What have we covered?

- Looked at how Amdahl’s Law can improve performance
- Looked at single and multithreaded programs
- Looked at some of the many Open Source Big Data tools that are available
- Looked at how and why some languages are better than others for a particular application

Next: Getting Twitter developer accounts
References I


References II


References III


How long is the coast of the Britain?

- Question raised by Richardson [9]
- Popularized by Mandelbrot [7]
- Foundational question in Chaos Theory [4]

Varies from $\approx 2,400$ to $\approx 3,400$ km depending on your yardstick[7]
Self referential curves

Curves that look like themselves.

- Richardson derived:
  \[ L(G) = MG^{1-D} \]
- It was ignored
- \( D \) is the dimensional characteristic [7]

Fig. 2. Nonrectifiable self-similar curves can be obtained as follows. Step 1: Choose any of the above drawings. Step 2: Replace each of its \( N \) legs by a curve deduced from the whole drawing through similarity of ratio \( 1/4 \). One is left with a curve made of \( N^2 \) legs of length \( (1/4)^p \). Step 3: Replace each leg by a curve obtained from the whole drawing through similarity of ratio \( (1/4)^q \). The desired self-similar curve is approached by an infinite sequence of these steps.
Simple algorithms yield things of beauty.
In 2 and 3D.

- Mandelbrot's equation:
  \[ z_{n+1} = z_n^2 + c \]
  where \( c \) is complex

- Mandelbrot curve is self referential
Big data problems are addressed in your computer.

With Koch and Mandelbrot, we were looking deeper and deeper. What happens if we go higher instead of deeper?

<table>
<thead>
<tr>
<th>Concept</th>
<th>Computer</th>
<th>Big Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paralizable</td>
<td>Cores</td>
<td>Processing nodes</td>
</tr>
<tr>
<td>Data locality</td>
<td>Cache (L1, L2, etc.)</td>
<td>HDFS</td>
</tr>
<tr>
<td>Coordination</td>
<td>OS</td>
<td>Hadoop</td>
</tr>
<tr>
<td>Output</td>
<td>RAM</td>
<td>HDFS</td>
</tr>
</tbody>
</table>

We will be bringing these ideas out into the open.