datanami
DATA SCIENCE • AI • ADVANCED ANALYTICS
(https://www.datanami.com)

Select Language ▼
Translation Disclaimer (/termsofuse.html#translation)
Search this site Search

Home (http://www.datanami.com/) About (https://www.datanami.com/about/)
Whitepapers (http://www.datanami.com/whitepaper/) Events (http://www.datanami.com/events/)
Subscribe (https://www.datanami.com/subscribe/)

Follow Datanami:
(http://www.facebook.com/pages/Datanami/124760547631
(http://www.twitter.com/datanami)
(http://www.linkedin.com/groups/Big-Data-News-Network-4166980)
(http://www.datanami.com/feed/)

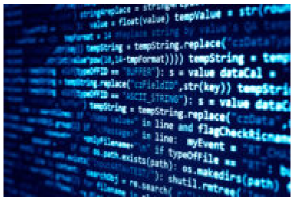
HOME (HTTP://WWW.DATANAMI.COM/) FEATURES SECTORS APPLICATIONS TECHNOLOGIES VENDORS
JOB BANK (HTTP://WWW.DATANAMI.COM/JOB-BANK/)

Top
Stories
On

February 12, 2018

# Which Programming Language Is Best for Big Data?

Alex Woodie



*(Vintage Tone/Shutterstock)*

Nothing is quite so personal for programmers as what language they use. Why a data scientist, engineer, or application developer picks one over the other has as much to do with personal preference and their employers' IT culture as it does the qualities and characteristics of the language itself. But when it comes to big data, there are some definite patterns that emerge.

The most important factor in choosing a programming language for a big data project is the goal at hand. If the organization is manipulating data, building analytics, and testing out machine learning models, they will probably choose a language that's best suited for that task. If the organization is looking to operationalize a big data or Internet of Things (IoT) application, there are another set of languages that excel at that.

In the data science exploration and development phase, the most popular language today unquestionably is Python. One big reason for Python's popularity is the plethora of tools and libraries available to help data scientists explore big data sets. Python was recently ranked the number one language by IEEE Spectrum (https://spectrum.ieee.org/computing/software/the-2017-top-programming-languages), where it moved up two spots to beat C, Java, and C++, although Python trails these languages on the TIOBE Index (https://www.tiobe.com/tiobe-index/). As a general purpose language, Python is also widely used outside of data science, which only adds to its usefulness.

Another popular data science language is R, which has long been a favorite of mathematicians, statisticians, and hard sciences. The SAS environment from the company of the same name (http://www.sas.com/) continues to be popular among business analysts, while MathWorks (http://www.matlab.com/)' MATLAB is also widely used for the exploration and discovery phase of big data. You also can't go far in data science without knowing some SQL, which remains a very useful language.

The choice of data science language may also be determined what notebook a data scientist is using.  Jupyter is the successor to the iPython notebook, and as such is closely aligned with Python, but it also supports R, Scala, and Julia. The Apache Zeppelin notebook includes Python, Scala, and SparkSQL support.

(https://2s7gjr373w3x22jf92z99mgm5w-wpengine.netdna-ssl.com/wp-

Programmers will often opt for a different set of languages when it comes to developing production analytics and IoT apps. While they may choose Python or R during the experimental phase of the project, programmers will often rewrite the application and re-implement the machine learning algorithms using entirely different languages.

Java continues to be a very popular choice owing to the large number of Java developers in the world, as well as the fact that some popular frameworks, such as Apache Hadoop, were developed in Java. Scala, which runs

This Just In | Most Read

March 15, 2018

Lexalytics Launches Support for Unicode Emoji 11.0 (https://www.datanami.com/this-just-in/lexalytics-launches-support-for-unicode-emoji-11-0/)
Proscia Inks Agreement to Bring Deep-Learning Technology to Dermatopathology (https://www.datanami.com/this-just-in/proscia-inks-agreement-to-bring-deep-learning-technology-to-dermatopathology/)
AI Expo Global Introduces Speaker Line-up

Visit additional Tabor Communications publications

(https://hpcwire.com/)(http://www.enterpriseai.news/)(https://www.hpcwire.jp/)

inside the Java Virtual Machine (JVM), is also widely used in data science; Apache Spark was written in Scala, and Apache Flink was written in a combination of Java and Scala.

content/uploads/2018/02/pro_lang_shutterstock_Alexander-Supertramp.jpg)

*There are many factors that go into choice of programming languages (Alexander Supertramp/Shutterstock)*

However, for some production applications, developers still favor lower-level languages that run closer to the iron. When speed and latency matter, many developers turn to C and C++ to get them what they want.

MapR Technologies developed its own big data platform, which contained a Hadoop runtime, a NoSQL database, and real-time streaming. But instead of writing its MapR-FS file system in Java, as HDFS was developed, it wrote it in C and C++. As MapR's Senior Staff Software Engineer Smidth Panchamia explained in this MapR blog post (https://mapr.com/blog/high-performance-c-apis-mapr-db/), it's tough to beat C and C++ for some tasks.

"Native languages like C/C++ provide a tighter control on memory and performance characteristics of the application than languages with automatic memory management," Panchamia writes. "A well written C++ program that has intimate knowledge of the memory access patterns and the architecture of the machine can run several times faster than a Java program that depends on garbage collection. For these reasons, many enterprise developers with massive scalability and performance requirements tend to use C/C++ in their server applications in comparison to Java."

(https://2s7gjr373w3x22jf92z99mgm5w-wpengine.netdna-ssl.com/wp-content/uploads/2014/12/java-8-logo.png)

Bloomberg (http://www.bloomberg.com/) uses Python for much of its data science exploratory work that goes into services delivered in the Bloomberg Terminal. But when it comes to writing the actual programs that feed data to customers in real time, it turned to C++.

"At the heart, it's a C++ shop," Bloomberg's Head of Data Science Gideon Mann told *Datanami* last year. "Most of the time, when we're doing data science, it's really to build machine learning products. And because we have all of these real time latency constraints, we don't want to use something like Python or Java, where you're going have garbage collection. You need to be a little worried about intermediate lag. By building out everything in C++, you can deploy it and have a fair amount of latency guarantees."

Another C++ aficionado is Dor Laor, CEO of ScyllaDB (http://www.scylladb.com/), which is a drop-in replacement for the Apache Cassandra NoSQL database. While Cassandra was written in Java, ScyllaDB was written in C++.

Laor, who also helped develop the KVM hypervisor, says lower-level languages in general are better for developing system software and databases. He points out that software giant Oracle (http://www.oracle.com/), which controls Java, opted to write its eponymous database in C. IBM (http://www.ibm.com/)'s DB2 was written in a combination of C and C++, he pointed out. "Even Mongo (http://www.mongodb.com/) is written in C++," he said.

By essentially rewriting Cassandra in C++ and avoiding the garbage collection associated with JVM, ScyllaDB is able to achieve orders-of-magnitude performance gains over Cassandra, Laor claimed.

"If you run Cassandra, then you need to reserve some amount [of memory] for Java," he tells *Datanami*. "And you also need to reserve additional amounts for off-heap data structures that are too heavy for Java too handle. And you also need to preserve enough memory for the Linux page cache to cache to disk. Forget about performance — just to tune it, it's a nightmare."

(https://2s7gjr373w3x22jf92z99mgm5w-wpengine.netdna-ssl.com/wp-content/uploads/2016/08/python_logo_1.png)

ScyllaDB was developed using C++ version 17. "It's the latest and greatest of C++, the cutting edge," Laor says. "It allows us to use really fancy language options, but it's also complex, so there's a big learning curve…even the time it takes you to compile the database is very long."

However, there are downsides to developing a database in C++, Laor admits. For starters, the increased complexity of the C++ source code means fewer developers will be able to contribute to the ScyllaDB project, which is open source. Plus, for some developers, letting the JVM handle memory gives them more time to develop better algorithms, which may be a good tradeoff.

The real-time stream analytics platform SQLstream (http://www.sqlstream.com/) was also developed in C++. "Not only do you get better performance from the code, but even more importantly, it's the lack of garbage collection," SQLstream CEO and founder Damian Black told *Datanami* last year.

**Performance with Industrial Internet of Things**

💬 NO COMMENTS

(http://tci.taborcommunications.com/…
11-14/65lrgg)

**Demand-Presenting Anzo Smart Data Lake®**

💬 NO COMMENTS

(http://tci.taborcommunications.com/…
11-03/6545vj)

## Contributors

Alex Woodie

Editor in Chief

George Leopold

Contributing Editor

Steve Conway

Hyperion Research

Tiffany Trader

Contributing Editor

## Featured Events

**Percona Live Open Source Database Conference (https://www.datanami.com/event/percona-live-open-source-database-conference-us/)**

April 23 - April 25

Santa Clara CA United States

**BigIT Diverse 2018 (https://www.datanami.com/even diverse-2018/)**

May 8 - May 9

**Artificial Intelligence in Automotive Conference (https://www.datanami.com/event/ai-automotive-conference/)**

May 17 - May 18

Silicon Valley CA United States

View More… (https://www.datanami.com/events/)

Managing the memory itself gives SQLstream a 5x performance boost over Java, Black says. “Not only that, we have lock-free execution, which is not easy to do,” he continued. “It’s a trendy thing but it’s really hard to do. You have to have a true declarative system, which we do have. We don’t transact any of the input streams or data or window objects, unlike almost any of the other streaming platforms.”

Since Apache Hadoop was written in Java, the developers at Hortonworks (http://www.hortonworks.com/) use Java for many of the sub-projects and other open source products that make up the Hortonworks Data Platform (HDP). It also programs in Java for Hortonworks Data Flow (HDF), which is based on the Java-based Apache NiFi. But for IoT apps, NiFi has a secret weapon: C++.

“NiFi has a pretty cool thing called MiniFi,” Hortonworks co-founder and Chief Product Officer Arun Murthy told *Datanami* last year. “It’s C++ driver you throw on cellphone or a security camera. So you can collect data from IoT-ish devices, all the way [out on the edge], secured and encrypted, and move it to your enterprise data center.”

(https://2s7gjr373w3x22jf92z99mgm5w-wpengine.netdna-ssl.com/wp-content/uploads/2018/02/C-logo.png)

Another streaming product based on C++ is the Concord framework that came out of the ad tech world. When YieldMo (wwww.yieldmo.com) had trouble getting Apache Storm (developed in Java and a JVM-compliant language called Clojure) to scale, a group of developers at the company, including Shinji Kim, decided to build their own real-time streaming system based on the MillWheel paper from Google (http://www.google.com/).

The resulting Concord product – which was acquired last fall (https://www.datanami.com/2016/10/04/acquisition-validates-concords-event-based-approach/) by Akamai Technologies (http://www.akamai.com/) – was written in C++ and implemented on the Mesos resource scheduler. “If you run that on Hadoop MapReduce jobs, if something fails, it definitely can cause a certain behavior, like cascading failure or a cluster-wide failure if one of your jobs doesn’t run well,” Kim told *Datanami*. “Or there could be an issue with the JVM where if you get high influx of traffic all of a sudden, if a GC [garbage collection] kicks in… there’s a lot of computations that you need get right.”

Before it was acquired by Apple (http://www.apple.com/) two years ago, Turi (formerly GraphLab and Dato) developed a popular machine learning framework that included graph algorithms. While the framework as a whole was open source and has Python APIs for data scientists to develop in, the underlying machine learning engine, based in C++, remained proprietary. There was good reason for that, as Turi’s Rajat Arya explained.

“Most academic papers and almost all vendors are talking about how long to train a model,” Arya told *Datanami*. “It turns out you really care about how long it takes to score a model or get a prediction. The real time prediction is what’s important because that’s what’s driving the business.”

By writing the engine in C++, Turi could be ensured a certain level of performance. “Open source is a great teaching tool. It gets a lot more people plugged in,” Arya said. “But the ability to get something done in a week is much more important. Open source can’t fill that gap.”

**Related Items:**

Data Science Ed: 5 Tips for Undergrads (https://www.datanami.com/2017/10/16/data-science-ed-5-tips-undergrads/)

Python Versus R in Apache Spark (https://www.datanami.com/2015/07/13/python-versus-r-in-apache-spark/)

Will Scala Take Over the Big Data World? (https://www.datanami.com/2015/08/10/will-scala-take-over-the-big-data-world/)

**Share this:**

Tweet Share Share G+

reddit this! (//www.reddit.com/submit?url=https://www.datanami.com/2018/02/12/programming-language-best-big-data/)

Tags: big data (https://www.datanami.com/tag/big-data/), C (https://www.datanami.com/tag/c/), Java (https://www.datanami.com/tag/java/), Julia (https://www.datanami.com/tag/julia/), matlab (https://www.datanami.com/tag/matlab/), programming languages (https://www.datanami.com/tag/programming-languages/), python (https://www.datanami.com/tag/python/), R (https://www.datanami.com/tag/r/), sas (https://www.datanami.com/tag/sas/), scala (https://www.datanami.com/tag/scala/)