DesignMind

Building Your Big Data Team

With all the buzz around Big Data, many companies have decided they need some sort of Big Data initiative in place to stay current with modern data management requirements. Platforms like Hadoop promise scalable data storage and processing at lower cost than traditional data platforms, but the newer technology requires new skill sets to plan, build and run a Big Data program. Companies not only have to determine the right Big Data technology strategy, they need to determine the right staffing strategy to be successful.

What comprises a great Big Data team? Luckily the answer isn't that far from what many companies already have in house managing their data warehouse or data services functions. The core skill sets and experience in the areas of system administration, database administration, database application architecture / engineering and data warehousing are all transferable. What usually makes someone able to successfully make the leap from, say, Oracle DBA to Hadoop administrator is their motivation to learn and more curiosity about what's happening under the covers when data are loaded or a query is run. Here are the key roles your (Hadoop) Big Data team needs to fill to be complete.

The first three roles are about building and operating a working Hadoop platform:

1. Data Center and OS Administrator

This person is the one who stands up machines, virtual or otherwise, and presents a fully functional system with working disks, networking, etc. They're usually also responsible for all the typical data center services (LDAP/AD, DNS, etc.) and know how to leverage vendor-specific tools (e.g., Kickstart and Cobbler, or vSphere) to provision large numbers of machines quickly. This role is becoming increasingly sophisticated and/or outsourced with the move to infrastructure as a service approaches to provisioning, with the DC role installing hardware and maintaining the provisioning system, and developers interfacing with that system rather than humans.

About DesignMind

DesignMind leverages products from technology partners, including Microsoft, Cloudera, Qubole, DataStax, Tableau, MapR, Hortonworks, Datameer, and Platfora, to build data intensive business applications. Headquartered in San Francisco, we help businesses leverage their data to gain competitive advantage.

While this role usually isn't (shouldn't be) part of the core Big Data team, it's mentioned here because it's absolutely critical to building a Big Data platform. Anyone architecting a new platform needs to partner closely with the data center team to specify requirements and validate the build out.

BACKGROUND/PREVIOUS ROLE: This role typically already exists, either in the organization or via a cloud hosted solution's web site.

TRAINING/RETOOLING: If you still run your own data center operations, definitely look at large scale virtualization and server management technologies that support self-service provisioning if you haven't already.

"Anyone architecting a new platform needs to partner closely with the data center team to specify requirements and validate the build out."

2. Hadoop Administrator

This person is responsible for set up, configuration, and ongoing operation of the software comprising your Hadoop stack. This person wears a lot of hats, and typically has a DevOps background where they're comfortable writing code and/or configuration with tools like Ansible to manage their Hadoop infrastructure. Skills in Linux and MySQL / PostgreSQL administration, security, high availability, disaster recovery and maintaining software across a cluster of machines is an absolute must. For the above reasons, many traditional DBAs aren't a good fit for this type of role even with training (but see below).

BACKGROUND/PREVIOUS ROLE: Linux compute / web cluster admin, possibly a DBA or storage admin.

TRAINING/RETOOLING: Hadoop admin class from Cloudera, Ansible training from ansible.com, plus lots of lab time to play with the tools.

3. HDFS/Hive/Impala/HBase/MapReduce/Spark Admin

Larger Big Data environments require specialization, as it's unusual to find an expert on all the different components - tuning Hive or Impala to perform at scale is very different from tuning HBase, and different from the skills required to have a working high availability implementation of Hive or HttpFS using HAProxy. Securing Hadoop properly can also require dedicated knowledge for each component and vendor, as Cloudera's Sentry (Hive / Impala / Search security) is a different animal than Hortonworks's Knox perimeter security. Some DBAs can successfully make the transition to this area if they understand the function of each component.

DesignMind

BACKGROUND/PREVIOUS ROLE: Linux cluster admin, MySQL / Postgres / Oracle DBA. See Gwen Shapira's excellent blog on Cloudera's site called "the Hadoop FAQ for Oracle DBAs."

TRAINING/RETOOLING: Hadoop admin and HBase classes from Cloudera, plus lots of lab time to play with all the various technologies.

The next four roles are concerned with getting data into Hadoop, doing something with it, and getting it back out:

4. ETL / Data Integration Developer

This role is similar to most data warehouse environments - get data from source systems of record (other databases, flat files, etc), transforming it into something useful and loading into a target system for querying by apps or users. The difference is the tool chain and interfaces the Hadoop platform provides, and the fact that Hadoop workloads are larger. Hadoop provides basic tools like Sqoop and Hive that any decent ETL developer that can write bash scripts and SQL can use, but understanding that best practice is really ELT (pushing the heavy lifting of transformation into Hadoop) and knowing which file formats to use for optimal Hive and Impala query performance are both the types of things that have to be learned.

A dangerous anti-pattern is to use generic DI tools / developers to load your Big Data platform. They usually don't do a good job producing high performance data warehouses with non-Hadoop DBs, and they won't do a good job here.

BACKGROUND/PREVIOUS ROLE: Data warehouse ETL developer with experience on high-end platforms like Netezza, Vertica, or Teradata.

TRAINING/RETOOLING: Cloudera Data Analyst class covering Hive, Pig and Impala, and possibly the Developer class.

"Focus on delivering a usable data product quickly, rather than modeling every last data source across the enterprise."

5. Data Architect

This role is typically responsible for data modeling, metadata, and data stewardship functions, and is important to keep your Big Data platform from devolving into a big mess of files and haphazardly managed tables. In a sense they own the data hosted on the platform, often have the most experience with the business domain and understand the data's meaning (and how to query it) more than anyone else except possibly the business analyst. The difference between a traditional data warehouse architect and this role is Hadoop's typical use in storing and processing unstructured data. Usually this is outside the traditional data architect's domain, and could include access logs and data from a message bus.

For less complex environments, this role is only a part-time job and is typically shared with the lead platform architect or business analyst. An important caveat for data architects and governance functions in general—assuming your Big Data platform is meant to be used for query and application workloads (vs. simple data archival), it's important to focus on delivering a usable data product quickly, rather than modeling every last data source across the enterprise and exactly how it will look as a finished product in Hadoop or Cassandra.

BACKGROUND/PREVIOUS ROLE: Data warehouse data architect.

TRAINING/RETOOLING: Cloudera Data Analyst class covering Hive, Pig and Impala.

DesignMind

6. Big Data Engineer / Architect

This developer is the one writing more complex data-driven applications that depend on core Hadoop applications like HBase, Spark or SolrCloud. These are serious software engineers developing products from their data for use cases like large scale web sites, search and real-time data processing that need the performance and scalability that a Big Data platform provides. They're usually well versed in the Java software development process and Hadoop toolchain, and are typically driving the requirements around what services the platform provides and how it performs.

For the Big Data architect role, all of the above applies, but this person is also responsible for specifying the entire solution that everyone else in the team is working to implement and run. They also have a greater understanding of and appreciation for problems with large data sets and distributed computing, and usually some system architecture skills to guide the build out of the Big Data ecosystem.

BACKGROUND/PREVIOUS ROLE: Database engineers would have been doing the same type of high performance database-backed application development, but may have been using Oracle or MySQL as a backend previously. Architects would have been leading the way in product development that depended on distributed computing and web scale systems.

TRAINING/RETOOLING: Cloudera's Data Analyst, Developer, Building Big Data Applications, and Spark classes.

DesignMind

7. Data Scientist

This role crosses boundaries between analyst and engineer, bringing together skills in math and statistics, machine learning, programming, and a wealth of experience / expertise in the business domain. All these combined allow a data scientist to search for insights in vast quantities of data stored in a data warehouse or Big Data platform, and perform basic research that often leads to the next data product for engineers to build. The key difference between traditional data mining experts and modern data scientists is the more extensive knowledge of tools and techniques for dealing with ever growing amounts of data. See our series of blog posts on how to hire a data scientist.

BACKGROUND/PREVIOUS ROLE: Data mining, statistician, applied mathematics.

TRAINING/RETOOLING: Cloudera's Intro to Data Science and Data Analyst classes.

"The key difference between traditional data mining experts and modern data scientists is the more extensive knowledge of tools and techniques for dealing with ever growing amounts of data."

The final roles are the traditional business-facing roles for data warehouse, BI and data services teams:

8. Data Analyst

This role is largely what you'd expect - using typical database client tools as the interface, they run queries against data warehouses, produce reports, and publish the results. The skills required to do this are a command of SQL, including dialects used by Hive and Impala, knowledge of data visualization, and understanding how to translate business questions into something a data warehouse can answer.

BACKGROUND/PREVIOUS ROLE: Data Analysts can luckily continue to use a lot of the same analytics tools and techniques they're used to, and benefit from improvements in both performance and functionality.

TRAINING/RETOOLING: Cloudera's Data Analyst class covering Hive, Pig and Impala.

9. Business Analyst

Like the data analyst, this role isn't too different from what already exists in most companies today. Gathering requirements from end users, describing use cases and specifying product behavior will never go away, and are even more important when the variety of data grows along with volume. What's most different is the tools involved, their capabilities, and the general approach. The business analyst needs to be sensitive to the difference in end-user tools with Hadoop, and determine training needed for users.

DesignMind

BACKGROUND/PREVIOUS ROLE: Business / system analyst for previous data products, like a data warehouse / BI solution, or traditional database-backed web application.

TRAINING/RETOOLING: Possibly Cloudera's Data Analyst class covering Hive, Pig and Impala, especially if this role will provide support to end users and analysts.

What's the Right Approach?

All of the above roles mention Hadoop and target data warehousing use cases, but many of the same guidelines apply if Apache Cassandra or similar NoSQL platform is being built out. And of course there are roles that are needed for any software or information management project – project managers, QA / QC, etc. For those roles the 1 day Cloudera Hadoop Essentials class might be the best intro.

So with the above roles defined, what's the right approach to building out a Big Data team? A Big Data program needs a core group of at least one each of an architect, admin, engineer and analyst to be successful, and for a team that small there's still a lot of cross over between roles. If there's an existing data services or data warehouse team taking on the Big Data problem then it's important to identify who will take on those roles and train them, or identify gaps that have to be filled by new hires. Of course the team needs to scale up as workload is added (users, apps and data), also accounting for existing needs.

Transforming into a Big Data capable organization is a challenging prospect, so it's also a good idea to start with smaller POCs and use a lab environment to test the waters. The time invested in training and building out the team will pay off when it's time to leverage your data at scale with the production platform, improving the odds of your Big Data program being a success.

Mark Kidwell specializes in designing, building, and deploying custom end-to-end Big Data solutions and data warehouses.

"The time invested in training and building out the team will pay off when it's time to leverage your data at scale with the production platform, improving the odds of your Big Data program being a success."