Data engineer: qualifications and skills

Jeff Leek

@jtleek

www.jtleek.com

What does a data engineer do?

- Build data infrastructure
- Manage data storage and use
- Implement production tools

https://www.facebook.com/careers/jobs/a0IA000000G3ISfMAN/

facebook Sign Up

Email or Phone Password

Keep me logged in Forgot your password?

Work at Facebook | Teams | Locations | University Students

IT & Security

Find your job Menlo Park

Data Engineer

(Menlo Park, CA)

Facebook was built to help people connect and share, and over the last decade our tools have played a critical part in changing how people around the world communicate with one another. With over a billion people using the service and more than fifty offices around the globe, a career at Facebook offers countless ways to make an impact in a fast growing organization.

At Facebook, we have many opportunities to work with data each and every day. How would you like to work on data AND build some of the tools that are critical to moving & transforming this data into valuable & insightful information? If so, this is the right job for you. Our Enterprise Data Warehouse team works very closely with all aspects of data, both internal and external. We are looking for a Data Engineer with the Software Engineering chops to not only build data pipelines to efficiently and reliably move data across systems, but also to build the next generation of data tools to enable us to take full advantage of this data. In this role, your work will broadly influence the company's data consumers and analysts. You will get the opportunity to work with focused and scaled objectives in a company that has some of the most challenging problems to tackle. This is a full-time position based in our office in Menlo Park.

Responsibilities

- Build data expertise and own data quality for the awesome pipelines you build
- Architect, build and launch new data models that provide intuitive analytics to your customers
- Design, build and launch extremely efficient & reliable data pipelines to move data (both large and small amounts) to our ridiculously large Data Warehouse
- Design and develop new systems and tools to enable folks to consume and understand data faster
- Use your expert coding skills across a number of languages from PHP, Python and JavaScript

Apply

Please limit to 3 applications.

Other positions in IT

AV/VC Deployment Engineer
(Menlo Park)

AV/VC Deployment Lead
(Menlo Park)

Application Engineer, PL/SQL and Java
(Menlo Park)

Application Product Manager, Fixed Assets
(Menlo Park)

BI Platform Engineer
(Menlo Park)

Data Engineer (People Analytics)
(Menlo Park)

IT Capacity Engineer
(Menlo Park)

IT Technical, ANZ
(Sydney)

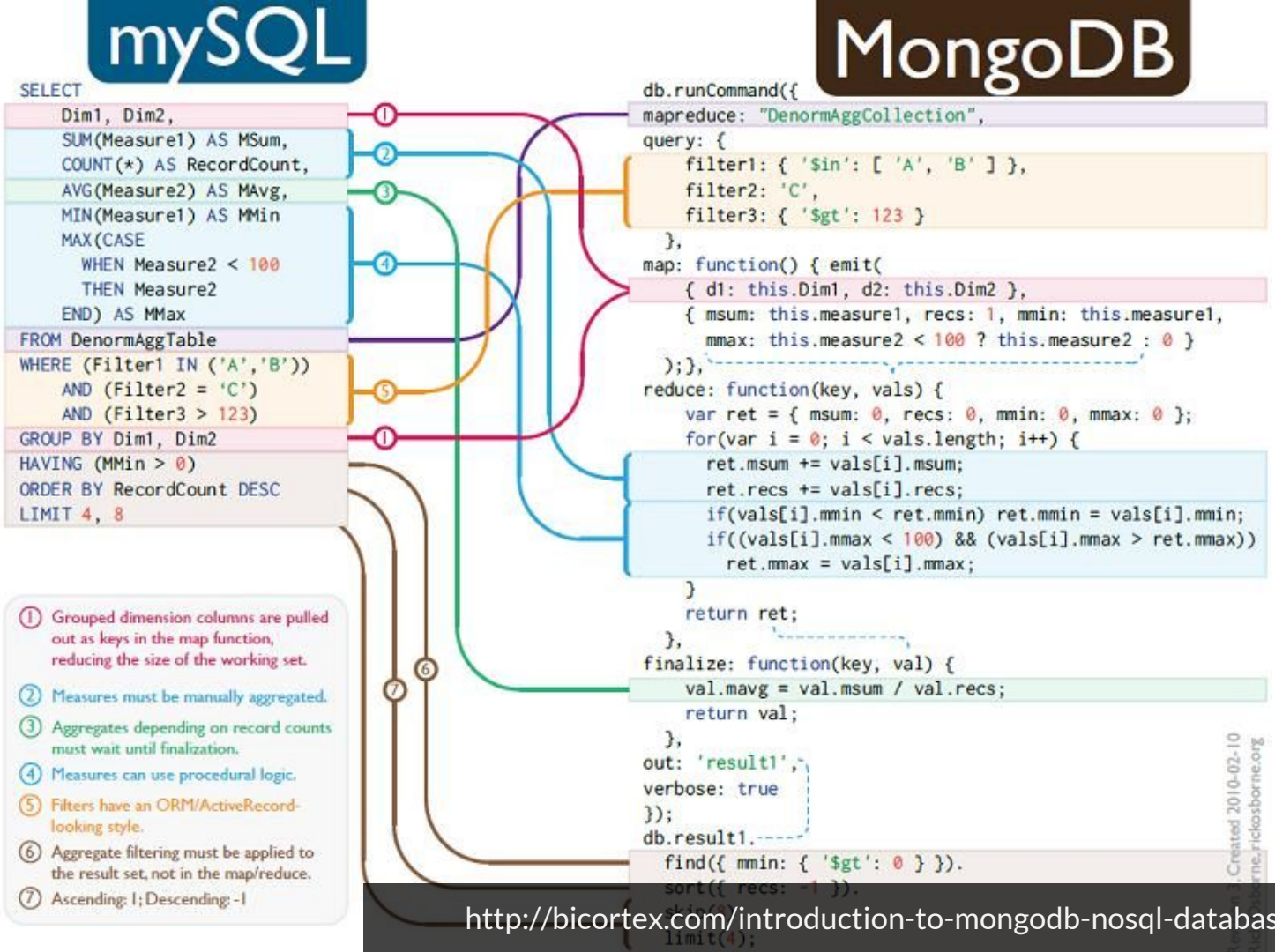
Lead Infrastructure Operations Engineer
(Menlo Park)

Lead Infrastructure Systems Engineer
(Menlo Park)

Back to all jobs

What skills do they need?

- Hardware knowledge
- Databases
- Data processing at scale
- Software engineering


mySQL
MongoDB
SELECT
Dim1, Dim2,
SUM(Measure1) AS MSum,
COUNT(*) AS RecordCount,
AVG(Measure2) AS MAvg,
MIN(Measure1) AS MMin
MAX(CASE
WHEN Measure2 < 100
THEN Measure2
END) AS MMax
FROM DenormAggTable
WHERE (Filter1 IN ('A','B'))
AND (Filter2 = 'C')
AND (Filter3 > 123)
GROUP BY Dim1, Dim2
HAVING (MMin > 0)
ORDER BY RecordCount DESC
LIMIT 4, 8
db.runCommand({
mapreduce: "DenormAggCollection",
query: {
filter1: { '$in': ['A', 'B'] },
filter2: 'C',
filter3: { '$gt': 123 }
},
map: function() { emit(
{ d1: this.Dim1, d2: this.Dim2 },
{ msum: this.measure1, recs: 1, mmin: this.measure1,
mmax: this.measure2 < 100 ? this.measure2 : 0 }
);},
reduce: function(key, vals) {
var ret = { msum: 0, recs: 0, mmin: 0, mmax: 0 };
for(var i = 0; i < vals.length; i++) {
ret.msum += vals[i].msum;
ret.recs += vals[i].recs;
if(vals[i].mmin < ret.mmin) ret.mmin = vals[i].mmin;
if((vals[i].mmax < 100) && (vals[i].mmax > ret.mmax))
ret.mmax = vals[i].mmax;
}
return ret;
},
finalize: function(key, val) {
val.mavg = val.msum / val.recs;
return val;
},
out: 'result1',
verbose: true
});
db.result1.
find({ mmin: { '$gt': 0 } }).
sort({ recs: -1 }).
limit(4);
1 Grouped dimension columns are pulled out as keys in the map function, reducing the size of the working set.
2 Measures must be manually aggregated.
3 Aggregates depending on record counts must wait until finalization.
4 Measures can use procedural logic.
5 Filters have an ORM/ActiveRecord-looking style.
6 Aggregate filtering must be applied to the result set, not in the map/reduce.
7 Ascending: 1; Descending: -1
Created 2010-02-10
http://bicortex.com/introduction-to-mongodb-nosql-database-for-sql-developers-part-3/

- Background of data engineers
  - Computer science and engineering
  - Quantitative + computer science
  - Information technology

https://www.cs.jhu.edu/

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

JUMP TO DEPARTMENT & PROGRAM WEBSITES

Department of Computer Science

ABOUT PEOPLE RESEARCH UNDERGRADUATE STUDIES GRADUATE STUDIES NEWS & EVENTS ALUMNI & GIVING IT SUPPORT

'Typosquatting': How 1 mistyped letter could lead to ID theft

Do you know what "typosquatting" is? CS Matt Green explains what to know to prevent identity theft.

GET THE STORY ›

DEPARTMENT NEWS ›

'Typosquatting': How 1 mistyped letter could lead to ID theft ›

Hopkins looks to code to identify a 'major and underappreciated' health problem, CS's Suchi Saria, The Baltimore Sun ›

Computer algorithm could aid in early detection of life-threatening sepsis ›

CS Summer Research Expeditions Program Provides Mentorships for Undergraduates ›

UPCOMING EVENTS

SEP 22 Tue — 10:30 am Annual CS State of the Departmen... @ Hackerman Hall Room B-17

Add ▾ View Calendar →

Solutions versus software

http://dataconomy.com/sql-vs-nosql-need-know/

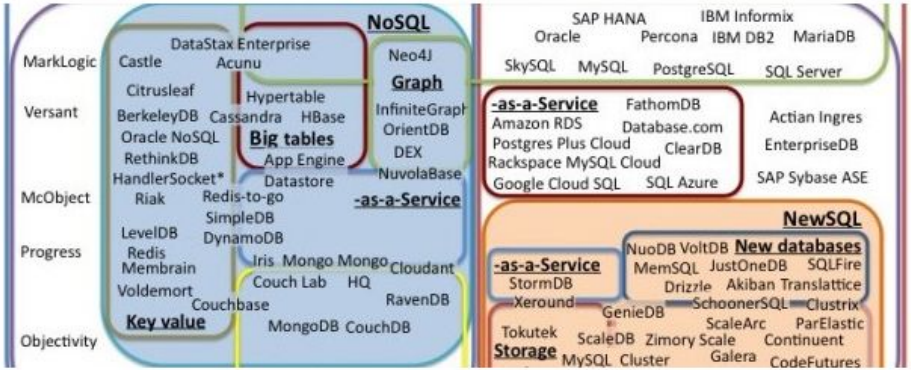
DATACONOMY

UNDERSTANDING BIG DATA

SQL VS. NOSQL- WHAT YOU NEED TO KNOW

EILEEN MCNULTY · JULY 1, 2014

16 COMMENTS ♥0 4.3K 1



8.2K FOLLOWERS

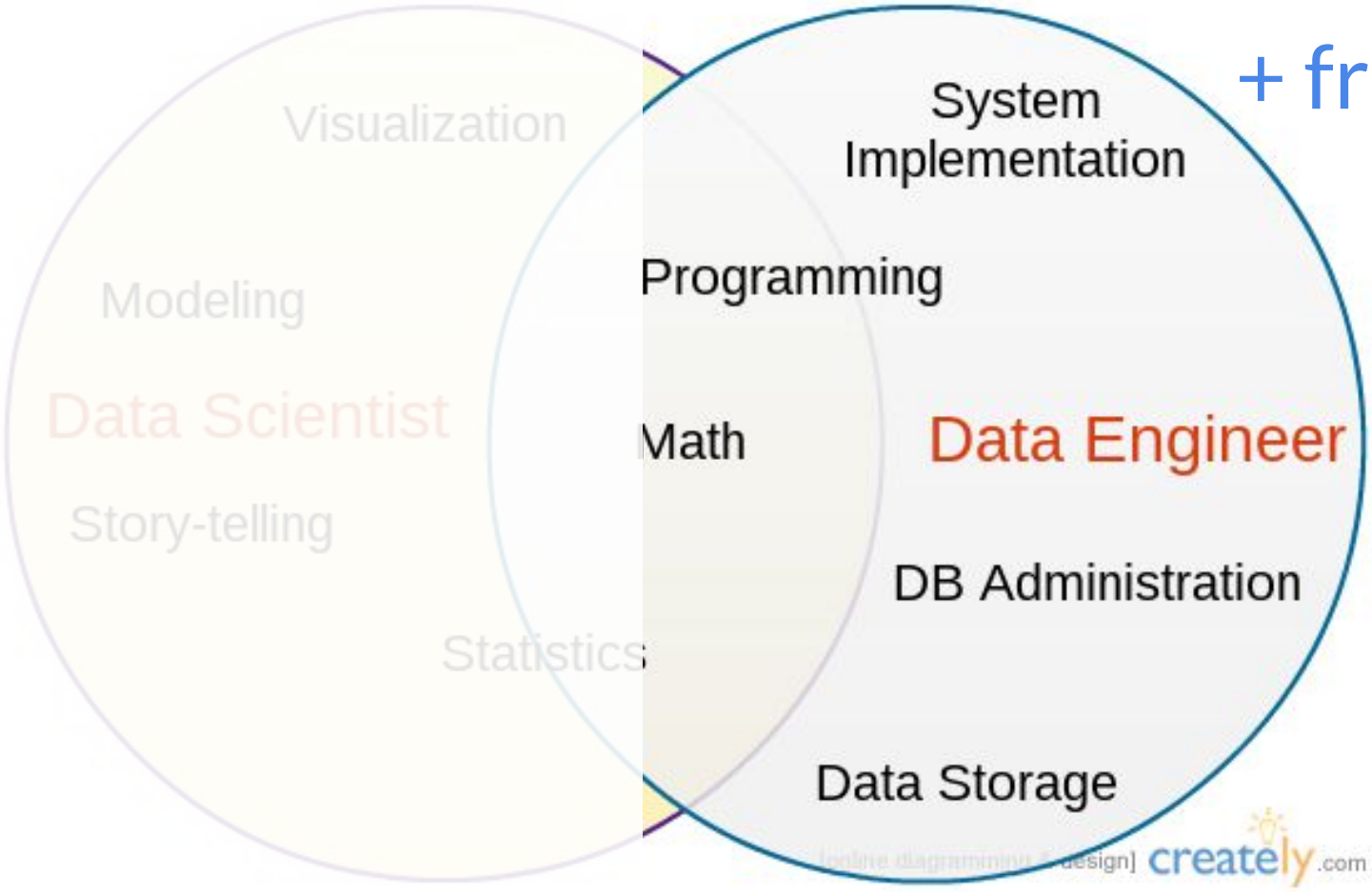
923 FANS

POPULAR RECENT COMMENTS

SQL VS. NOSQL- WHAT YOU NEED TO KNOW

Key characteristics

- Willing to find answers on their own
- Knows a bit of data science
- Works well under pressure
- Friendly but relentless


Visualization
Modeling
Data Scientist
Story-telling
Statistics
System Implementation
Programming
Math
Data Engineer
DB Administration
Data Storage
creately.com

+ friendly!