Big Data Engineering

with Python

Kholed Langsari

AI/ML Engineer, Software Architect,
Instructor at Fatoni University

langsari@ftu.ac.th

Big Data Engineering

Big Data Engineering

Learn to design data models, build data warehouses and data lakes, automate data pipelines, and work with massive datasets. At the end of the program, you'll combine your new skills by completing a capstone project.

Course Modules

1. Introduction
2. Python for Data Engineering
3. Data Modeling
4. Data Warehouses
5. Spark and Data Lakes
6. Automate Data Pipelines

Introduction

Introduction to Data Engineering

What does it mean to be a Data Engineer?

Setup works environment

Checking, installing and config

Python for Data Engineering

Python for Data Engineering

Essential Python language

Data Modeling

Introduction to Data Modeling

Understand the purpose of data modeling
Identify the strengths and weaknesses of different types of databases and data storage techniques
Create a table in Postgres and Apache Cassandra

Relational Data Models

Understand when to use a relational database
Understand the difference between OLAP and OLTP databases
Create normalized data tables
Implement denormalized schemas (e.g. STAR, Snowflake)

NoSQL Data Models

Understand when to use NoSQL databases and how they differ from relational databases
Select the appropriate primary key and clustering columns for a given use case
Create a NoSQL database in Apache Cassandran

Data Warehouses

Introduction to the Data Warehouses

Understand Data Warehousing architecture
Run an ETL process to denormalize a database (3NF to Star)
Create an OLAP cube from facts and dimensions
Compare columnar vs. row oriented approaches

Introduction to the Cloud with AWS

Understand cloud computing
Create an AWS account and understand their services
Set up Amazon S3, IAM, VPC, EC2, RDS PostgreSQ

Implementing Data Warehouses on AWS

Identify components of the Redshift architecture
Run ETL process to extract data from S3 into Redshift
Set up AWS infrastructure using Infrastructure as Code(IaC)
Design an optimized table by selecting the appropriate distribution style and sorting key

Data Lakes with Spark

The Power of Spark

Understand the big data ecosystem
Understand when to use Spark and when not to use it

Data Wrangling with Spark

Manipulate data with SparkSQL and Spark Dataframes
Use Spark for ETL purposes

Introduction to Data Lakes

Understand the purpose and evolution of data lakes
Implement data lakes on Amazon S3, EMR, Athena, and Amazon Glue
Use Spark to run ELT processes and analytics on data of diverse sources, structures, and vintages
Understand the components and issues of data lakes

Automate Data Pipelines

Data Pipelines

Create data pipelines with Apache Airflow
Set up task dependencies
Create data connections using hooks

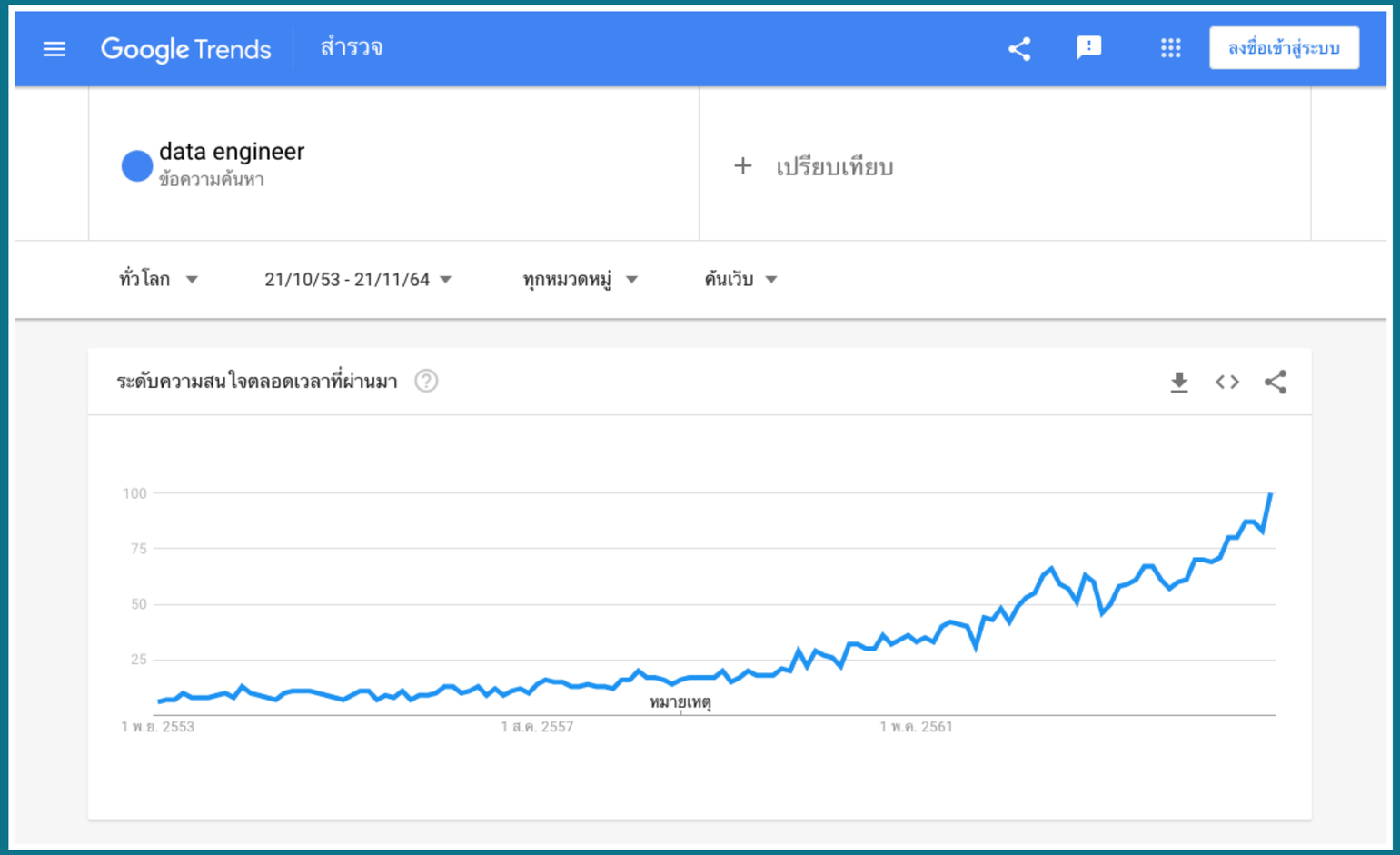
Data Quality

Track data lineage
Set up data pipeline schedules
Partition data to optimize pipelines
Write tests to ensure data quality
Backfill data

Production Data Pipelines

Build reusable and maintainable pipelines
Build your own Apache Airflow plugins
Implement subDAGs
Set up task boundaries
Monitor data pipelines

Big Data Engineering Trends

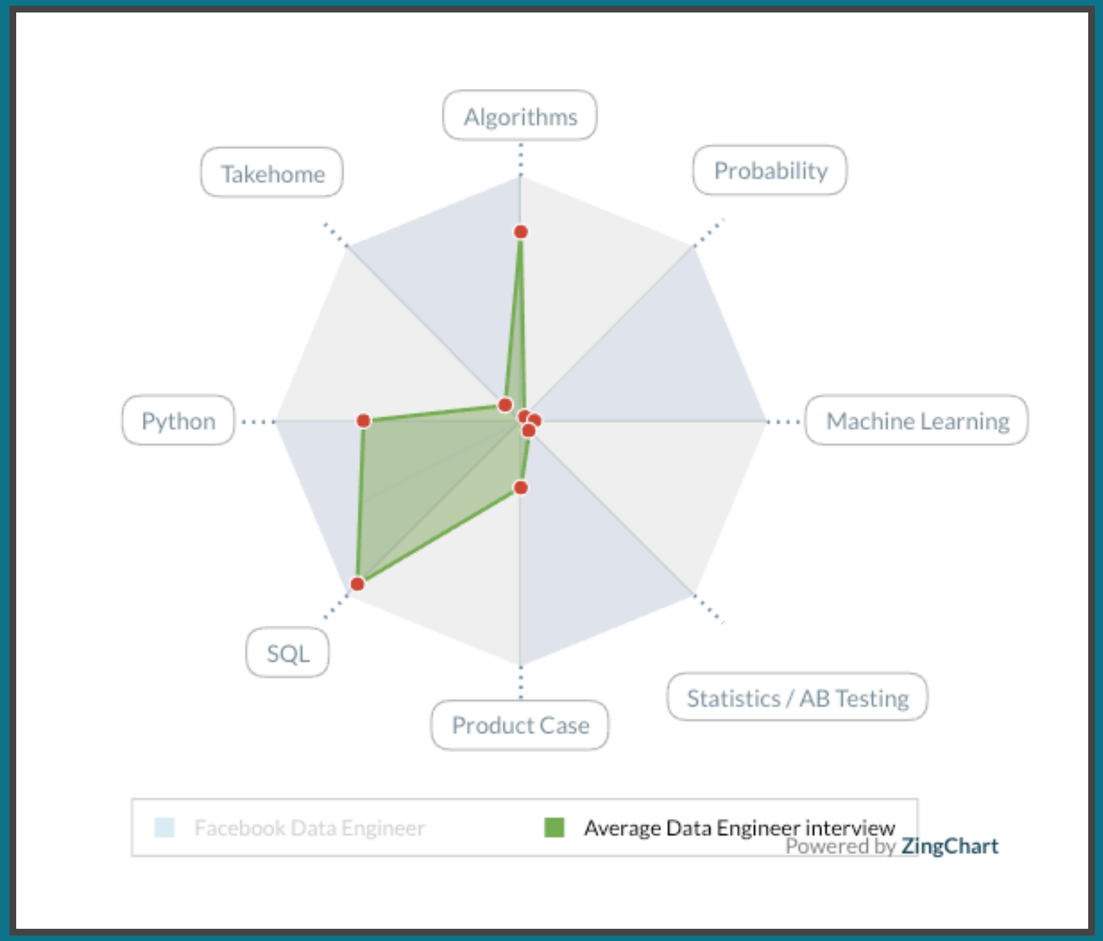


Data Engineer

Google Trends 2010-2021 (retrieved 7 Nov 2021)

“Data engineering is the new data science”

–Interview Query

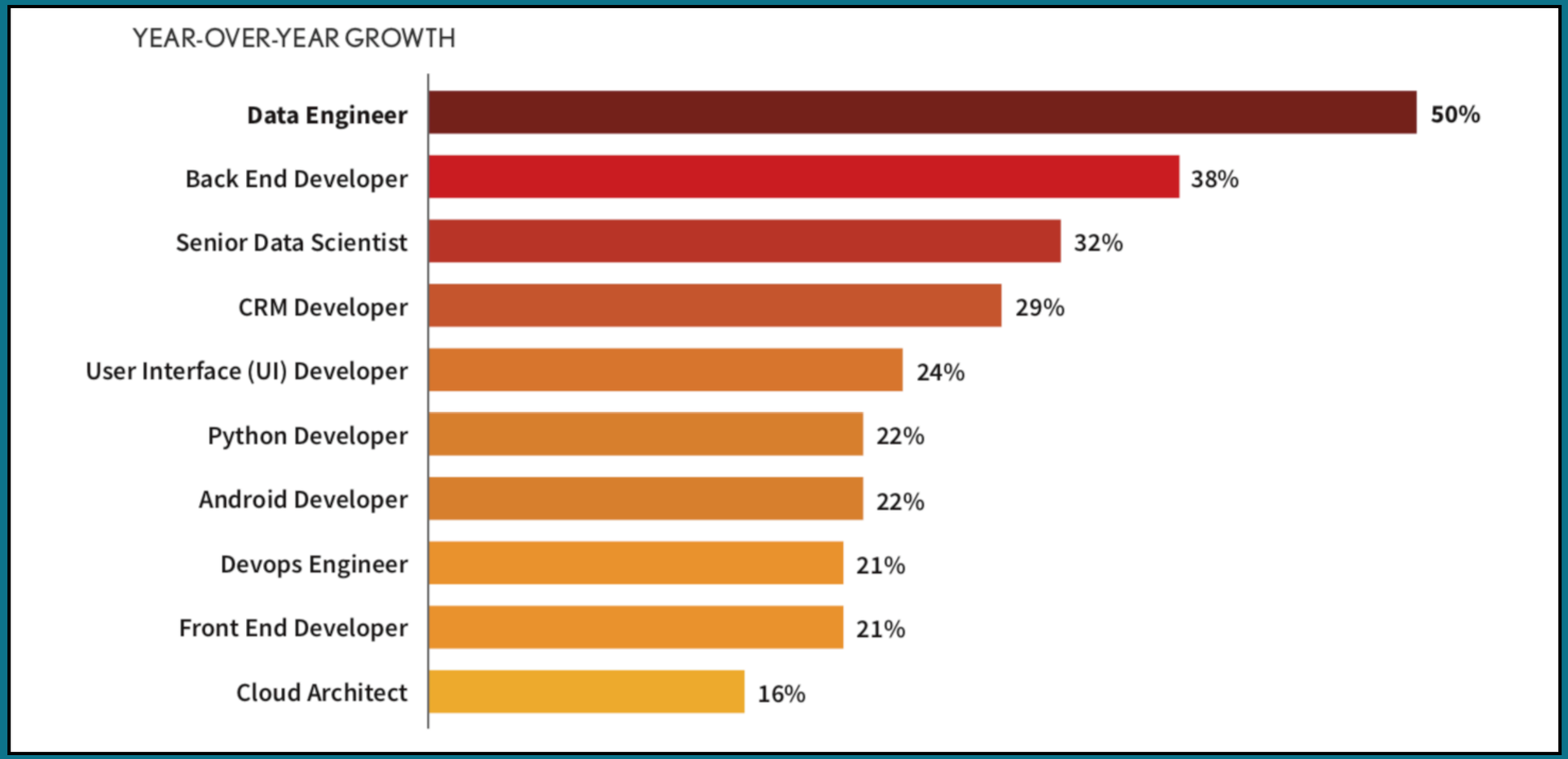


Data engineering interviews in the past year have grown by **40%**!

Data Engineer Radar Chart, Interview Query (2020)

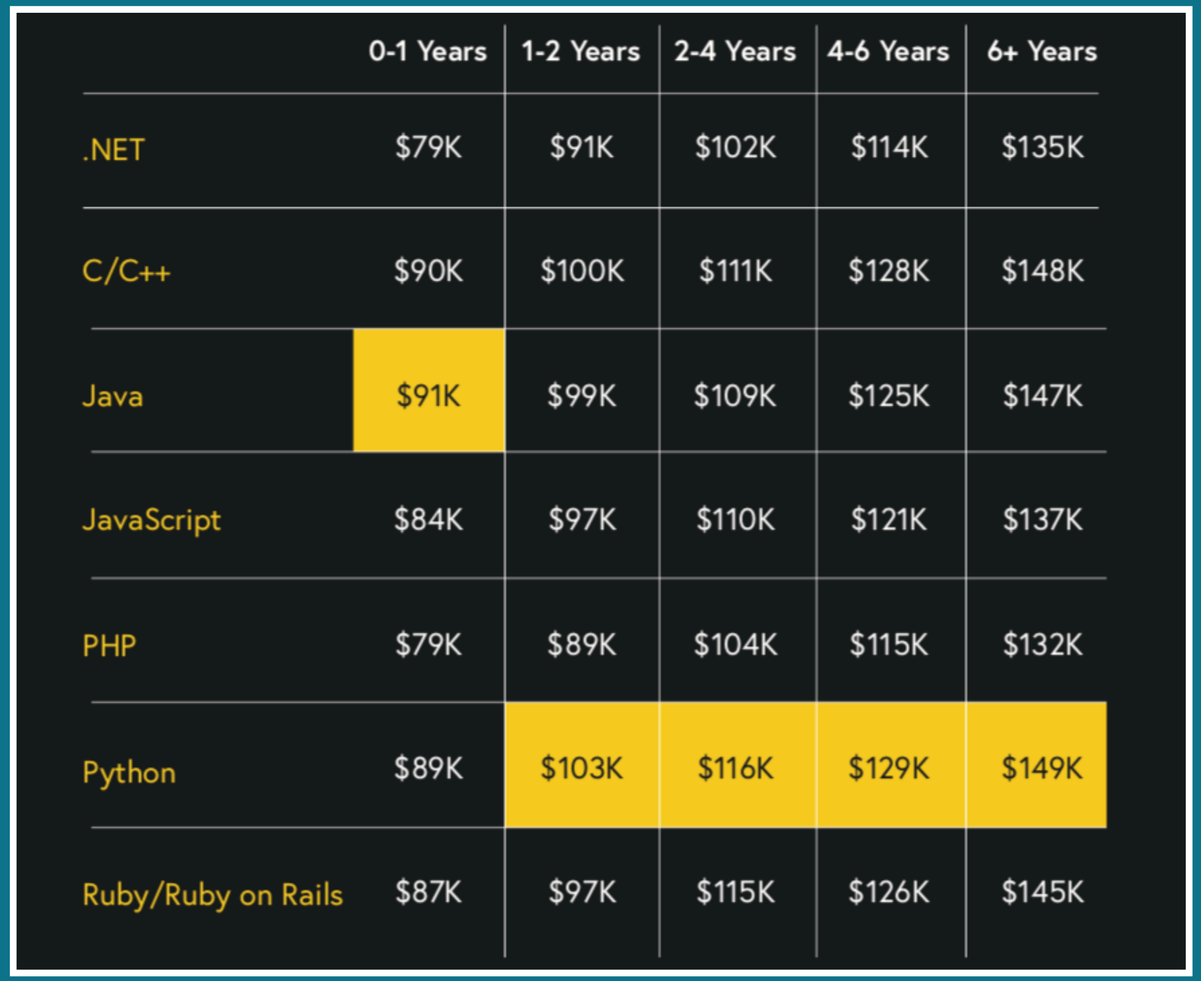
Data Engineering in demand

- Data Engineer is the **fastest-growing job** by **50%** YoY. Data Scientist growing by **32%** YoY (2019)
- Data Science interviews grew by **10%** compared to **Data Engineering** interviews which grew by **40%** in 2020 (Interview Query).
- **1,358 jobs** available (JobDB)
- **$112,493/yr** | Data Science ($117,212/yr) | data analyst ($68,000) | database administrator ($81,444) (Glassdoor)



Fastest Growing Tech Occupations

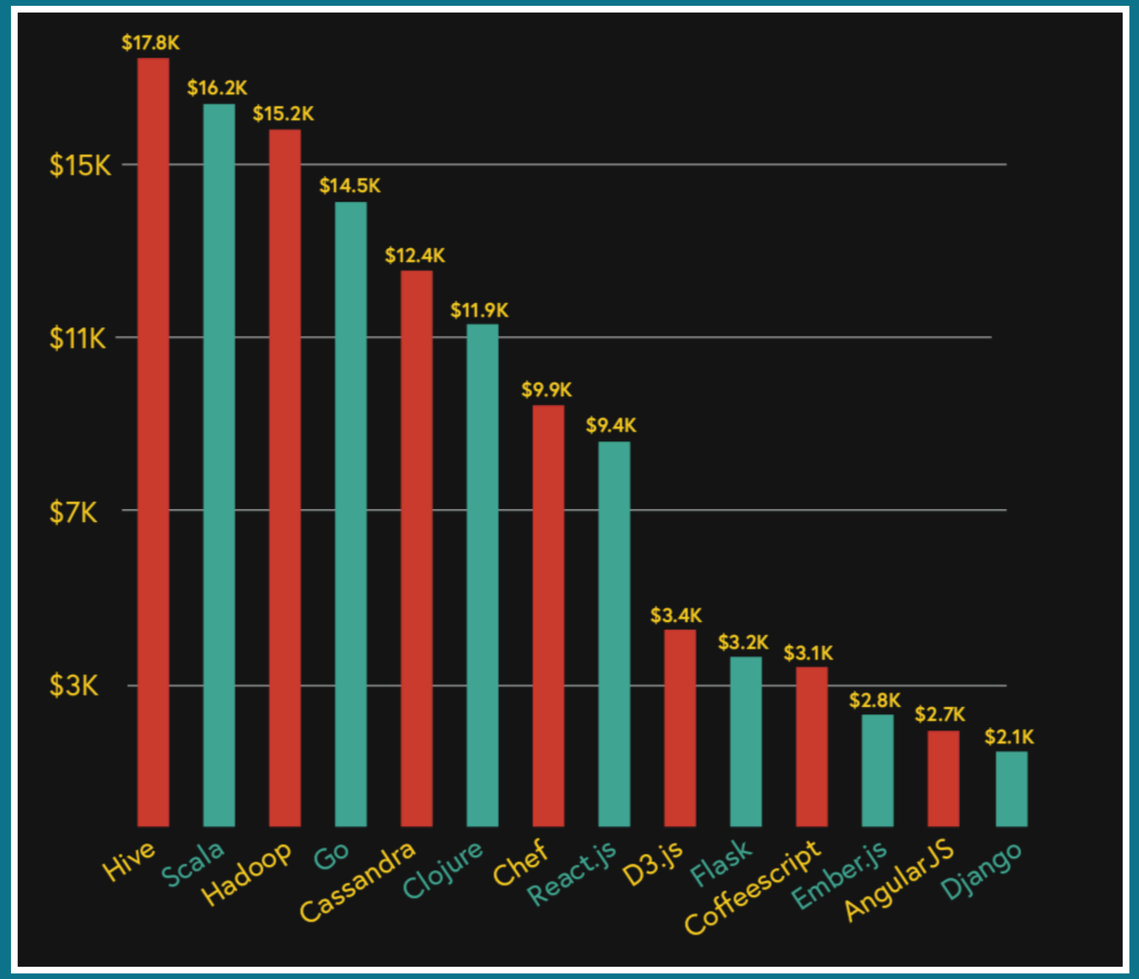
DICE's Tech Job Report, 2020



	0-1 Years	1-2 Years	2-4 Years	4-6 Years	6+ Years
.NET	$79K	$91K	$102K	$114K	$135K
C/C++	$90K	$100K	$111K	$128K	$148K
Java	$91K	$99K	$109K	$125K	$147K
JavaScript	$84K	$97K	$110K	$121K	$137K
PHP	$79K	$89K	$104K	$115K	$132K
Python	$89K	$103K	$116K	$129K	$149K
Ruby/Ruby on Rails	$87K	$97K	$115K	$126K	$145K

Popular Programming Skills

Vettery Salary Report 2020



Specialty Tech Programming Skills

Vettery Salary Report 2020

Weekly Open Questions

Reply through @thefutureisdata

LANG
SA
RI

วิศวกรรมข้อมูลขนาดใหญ่คืออะไร?

What is big data engineering?

Big Data Engineering

with Python

Kholed Langsari

AI/ML Engineer, Software Architect,
Instructor at Fatoni University

langsari@ftu.ac.th