Zara Blackwood

Big Data Executive

Phone: **(415) 555-0198**
Address: **San Francisco, CA**
Website: **https://linkedin.com/in/zarablackwood**
Email: **zara.blackwood@email.com**

- Big Data Engineer with 2+ years designing scalable data pipelines using Apache Spark 3.2+, Hadoop 3.3, and AWS EMR
- Reduced data processing time by 65% through optimization of ETL workflows handling 15TB+ daily transaction data
- Built real-time streaming applications using Kafka and Flink processing 200K+ events per second with 99.9% uptime
- Experienced in distributed computing frameworks, cloud-native architectures, and performance optimization for petabyte-scale datasets

WORK EXPERIENCE

StreamTech Solutions

June 2022 - Present

Junior Data Engineer

- Architected distributed data processing pipeline using Apache Spark 3.2+ on Kubernetes, reducing batch processing time from 6 hours to 2.1 hours for 800GB daily customer behavior data
- Implemented real-time anomaly detection system using Kafka Streams and Cassandra, identifying fraudulent transactions with 96% accuracy and reducing false positives by 30%
- Optimized Hive queries and Spark configurations resulting in 75% reduction in cluster resource usage and $22K monthly AWS cost savings
- Built automated data quality framework using Great Expectations, reducing data incidents by 85% and improving downstream analytics reliability
- Mentored 2 junior engineers on Spark optimization techniques, distributed computing best practices, and debugging failed MapReduce jobs

DataFlow Innovations

January 2022 - May 2022

Data Analytics Intern

- Developed ETL pipelines using PySpark processing 50GB+ of IoT sensor data daily, improving data ingestion speed by 40%
- Created real-time dashboard using Kafka Connect and Elasticsearch, enabling stakeholders to monitor 10K+ connected devices
- Implemented data partitioning strategies in HDFS reducing query response times from 45 seconds to 8 seconds for analytical workloads
- Collaborated with senior engineers to migrate legacy batch processes to streaming architecture using Apache Flink

Stanford University

September 2021 - December 2021

Research Assistant - Distributed Computing Lab

- Built distributed machine learning pipeline processing 2TB+ of social media data using Spark MLlib and custom partitioning strategies
- Optimized memory management for large-scale graph processing, achieving 3x performance improvement on 100M+ node networks
- Published research on stream processing optimization techniques, contributing to Apache Beam Python SDK performance improvements

TECHNICAL SKILLS

Big Data Processing: Apache Spark 3.2+ (PySpark, Spark SQL, Spark Streaming, MLlib), Hadoop Ecosystem (HDFS, YARN, MapReduce), Apache Kafka, Apache Flink, Apache Beam

Programming Languages: Python (pandas, NumPy, scikit-learn, PySpark), Scala, Java, SQL, R

Data Storage & Databases: HDFS, Amazon S3, Azure Data Lake Storage, Apache Cassandra, MongoDB, HBase, PostgreSQL, Delta Lake, Apache Iceberg

Cloud Platforms: AWS (EMR 6.5+, Glue, Athena, Kinesis, S3), Azure (Databricks, Synapse Analytics, Event Hubs), GCP (Dataflow, BigQuery, Pub/Sub)

Data Processing & Analytics: Apache Hive, Presto, Apache Airflow, Databricks, Jupyter Notebooks, Apache Zeppelin

DevOps & Monitoring: Docker, Kubernetes, Apache Kafka Manager, Grafana, Prometheus, ELK Stack (Elasticsearch, Logstash, Kibana)

Technical Leadership: Data architecture design, technical documentation, cross-functional collaboration with data science and business intelligence teams, performance optimization, scalability planning for 10x data growth scenarios

EDUCATION

Stanford University

May 2022

Master of Science in Computer Science - Data Systems Track

Relevant Coursework: Distributed Computing, Large-Scale Data Mining, NoSQL Database Systems, Stream Processing, Machine Learning at Scale, Statistical Learning Theory

Thesis: "Optimizing Apache Spark Performance for Real-Time Graph Analytics on Petabyte-Scale Social Networks"

Key Projects:

- Built distributed recommendation engine processing 5TB+ of user interaction data using collaborative filtering on Spark, achieving 94% accuracy
- Developed real-time fraud detection pipeline using Kafka and Flink, processing 150K transactions/second with sub-100ms latency

University of California, Berkeley

May 2020

Bachelor of Science in Applied Mathematics | Data Science Concentration

Magna Cum Laude, GPA: 3.8/4.0

Key Projects:

- Built distributed image classification system processing 8TB dataset using PySpark and TensorFlow, achieving 97% accuracy on 10M+ images
- Developed streaming analytics platform using Kafka Streams analyzing Twitter data at 50K tweets/second for sentiment analysis

PROFESSIONAL CERTIFICATIONS

AWS Certified Big Data - Specialty

2023

Amazon Web Services

Databricks Certified Associate Developer for Apache Spark 3.0

2023

Databricks

Google Cloud Professional Data Engineer

2022

Google Cloud

Cloudera Certified Professional: Data Engineer

2022

Cloudera

AWARDS & RECOGNITION

2nd Place, Netflix Big Data Challenge 2023

2023

Netflix

- Developed distributed recommendation algorithm processing 200M+ user interactions using ensemble methods on Spark
- Improved prediction accuracy by 28% over baseline models while reducing training time by 45%

Winner, Stanford Data Mining Competition 2022

2022

Stanford University

- Analyzed 3TB+ of IoT sensor data to predict equipment failures using time series analysis and distributed computing
- Solution implemented by campus facilities management, preventing $75K in equipment downtime annually

Outstanding Graduate Student Award - Computer Science Department

2022

Stanford University

- Recognized for thesis work on Spark optimization and contributions to distributed systems research

PUBLICATIONS & TECHNICAL WRITING

Optimizing Spark Performance for Time Series Analysis at Scale

DataEngineering.io

Featured article with 8,000+ views, includes working code examples and performance benchmarks

Real-time Stream Processing: Kafka vs. Pulsar Performance Comparison

Medium - Towards Data Science

Implemented comprehensive benchmarks processing 1M+ events/second, 12K+ claps

Open Source Contribution: Apache Beam Python SDK

Apache Foundation

Merged pull request improving windowing function performance by 20% for streaming workloads

FEATURED GITHUB PROJECTS

spark-optimization-toolkit

https://github.com/zarablackwood

Custom Spark transformations and utilities reducing shuffle operations by 45% and memory usage by 30%

real-time-anomaly-detector

https://github.com/zarablackwood

Kafka Streams application processing 100K+ events/second with machine learning-based anomaly detection

data-quality-framework

https://github.com/zarablackwood

Automated testing and validation suite for petabyte-scale data pipelines with configurable quality rules

TECHNICAL PRESENTATIONS

Building a 12TB/day Analytics Pipeline on a Startup Budget

2023

PyCon 2023 Lightning Talk

- Presented cost-optimization strategies reducing cloud spend by 65% using spot instances and auto-scaling
- Slides downloaded 750+ times, implementation adopted by 5+ startups