

 PERSONUPPGIFTER

@ samira.haddad@example.se

+46 70 220548

Norrköping

 PROJEKT

Högupplöst väderhistorik

Dataprodukt som stödjer PySpark, batchjobb och ML-underlag, med förvaltade källor, versionshanterade transformationer oc...

 FÄRDIGHETER

Plattformer och programmering

PySpark · Delta Lake · Spark SQL · Airflow · Great Expectations

Kvalitet och drift

Datamodellering · Testning · CI/CD · Övervakning · Dokumentation

Samarbete

Kravinsamling · Kodgranskning · Incidenthantering · Tvärfunktionellt arbete

 CERTIFIERINGAR

Databricks Certified Associate Developer for Apache Spark

Certifieringsorganisation

Workshop i datakvalitet: Högupplöst väderhistorik

Svenska Data Engineering-nätverket

 UTBILDNING

Linköpings universitet

Masterexamen i statistik och maskininlärning

09/2014 - 06/2018

SPRÅK

Svenska

Engelska

Arabiska

C1

C1

Modersmål

SAMIRA HADDAD

PYSPARK DATA ENGINEER

 SAMMANFATTNING

PySpark-specialist som förbereder väder- och sensordata för prognosmodeller. Optimerar distribution och kvalitet tillsammans med meteorologer och ML-utvecklare.

 ERFARENHET

Östgöta Väderdata

PySpark Data Engineer

Norrköping

01/2022 - Nuvarande

- Optimerade PySpark-joiner för 3,7 miljarder väderrader; med två meteorologer sjönk nattkörningen från 4 timmar 10 minuter till 76 minuter.
- Bytte till Delta Lake och Z-order för 18 års stationshistorik; prognosteamet tog fram regionala urval 62 procent snabbare.
- La in Great Expectations-kontroller för 47 mätfält från 680 stationer; felaktig luftfuktighet nådde inte träningen i 11 releaser.
- Justerade autoskalning efter belastning i 9 Spark-jobb; månadskostnaden minskade 22 procent medan leveranstiden höll sig under 35 minuter.

Klimatflöde Norden

PySpark Data Engineer

Norrköping

04/2018 - 12/2021

- Parallelliserade inläsning av 2 800 väderfiler per dygn; klimatforskare fick validerat underlag före klockan 07.00 i 98 procent av fallen.
- Jämförde partitionering efter station och datum med ett team på 4; småfilsläsning gick från 38 till 12 minuter.
- Återskapade saknade observationer med versionshanterade regler; 91 procent av 26 000 luckor märktes med metod och osäkerhet.
- Dokumenterade 34 Spark-parametrar i Airflow; nya modellerare satte upp en reproducerbar körning på en dag i stället för tre.