

INTISARI

IMPLEMENTASI AIRFLOW DAN APACHE SPARK DALAM PERANCANGAN *DATA PIPELINE* UNTUK *DATA PROCESSING AUTOMATION* PADA *CLOUD-BASED SERVER*

Gigih Haryo Yudhanto

19/447091/SV/16810

Volume data yang besar seringkali tidak diimbangi dengan kualitas data yang memadai dan membuat informasi yang didapat menjadi bias sehingga rawan terjadi kesalahan dalam pengambilan keputusan. Berdasarkan permasalahan tersebut, penelitian dilakukan untuk membuat sistem pemrosesan data secara otomatis untuk meningkatkan kualitas data pada *data pipeline*. Sistem dibuat dengan memanfaatkan Apache Airflow sebagai *workflow orchestrator* dan *scheduler* untuk menjalankan *data pipeline* mulai dari *crawling* data, penyimpanan data, dan pemrosesan data. Data hasil *crawling* akan diproses oleh Apache Spark untuk meningkatkan kualitas data sebelum akhirnya divisualisasikan oleh Looker Datastudio. Dalam penelitian ini, dilakukan pengujian performa Apache Spark untuk mengetahui kecepatan pemrosesan data yang dilakukan. Dari hasil pengujian performa Apache Spark, menunjukkan variasi jumlah *core* pada *spark executor* memiliki pengaruh signifikan terhadap kecepatan pemrosesan data yang dilakukan, sementara variasi besar *memory* pada *spark executor* tidak memiliki pengaruh signifikan terhadap kecepatan pemrosesan data pada *spark executor* dengan jumlah *core* lebih dari satu. Selain itu, dilakukan juga pengujian kualitas data yang meliputi tiga parameter kualitas data yaitu *completeness*, *uniqueness*, dan *validity*. Berdasarkan hasil pengujian kualitas data, diketahui pemrosesan data dapat mengurangi *error* data pada pengujian *completeness* menjadi 0%, pengujian *uniqueness* menjadi 0%, dan pengujian *validity* menjadi 0%.

Kata kunci: kualitas data, pemrosesan data, Apache Airflow, Apache Spark