Dataflow for Real-time ETL/Integration

We're losing customers because we can't keep up with the pace of the business!

Trust me, I know! Our dashboards are only updated monthly.

Imagine the insight's we'd gain if we analyzed data in real-time with Dataflow!

Here's how Dataflow can help your business...

Faster Decision Making

Operational Efficiency

Smarter ML models

Fresher Dashboards

Better Data

Superior Customer Experiences

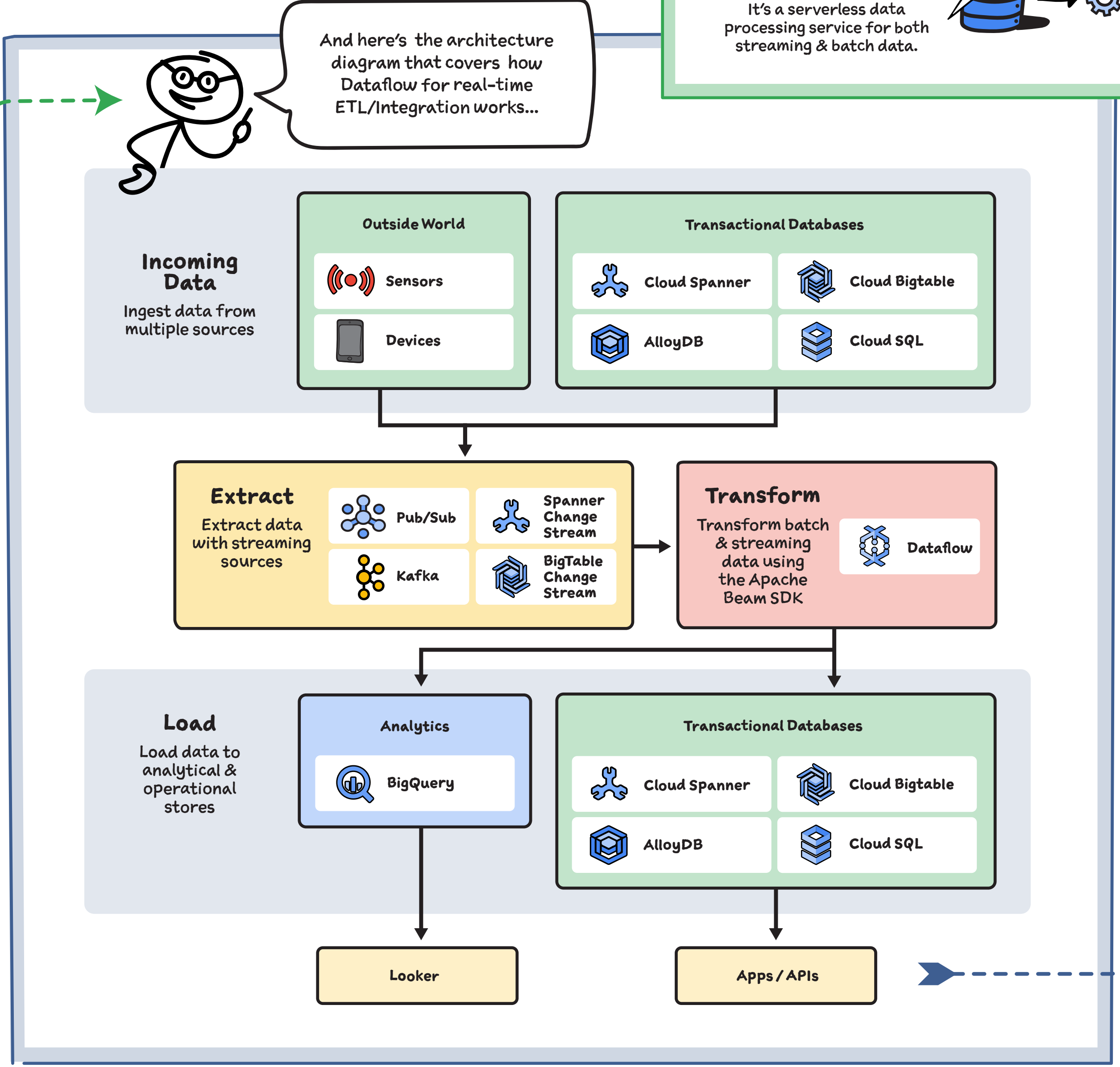Take a look at these technical benefits. Sweet!

Batch Stream

Unified codebase for batch and streaming pipelines with Apache Beam

Massive scalability to handle batch and streaming workloads with ease

Cost-effective processing with automatic resource optimization

Simplified operations with built-in fault tolerance for your data

Comprehensive monitoring for full visibility into pipeline health & performance

And here's the architecture diagram that covers how Dataflow for real-time ETL/Integration works...

What is Dataflow?

It's a serverless data processing service for both streaming & batch data.

Batch Data

Streaming Data

Incoming Data
Ingest data from multiple sources

Outside World
- Sensors
- Devices

Transactional Databases
- Cloud Spanner
- Cloud Bigtable
- AlloyDB
- Cloud SQL

Extract
Extract data with streaming sources
- Pub/Sub
- Spanner Change Stream
- Kafka
- BigTable Change Stream

Transform
Transform batch & streaming data using the Apache Beam SDK
- Dataflow

Load
Load data to analytical & operational stores

Analytics
- BigQuery

Transactional Databases
- Cloud Spanner
- Cloud Bigtable
- AlloyDB
- Cloud SQL

Looker

Apps / APIs

Keep in mind...

If deduplication is not needed, enable at-least-once mode for lower costs & faster processing

Leverage off-the-shelf Beam I/O connectors and Dataflow managed connectors to simplify code & minimize costs

Structure your data with Beam schemas for seamless integration across different databases

Weigh tradeoffs between latency & completeness of your data when selecting window configuration