SDF

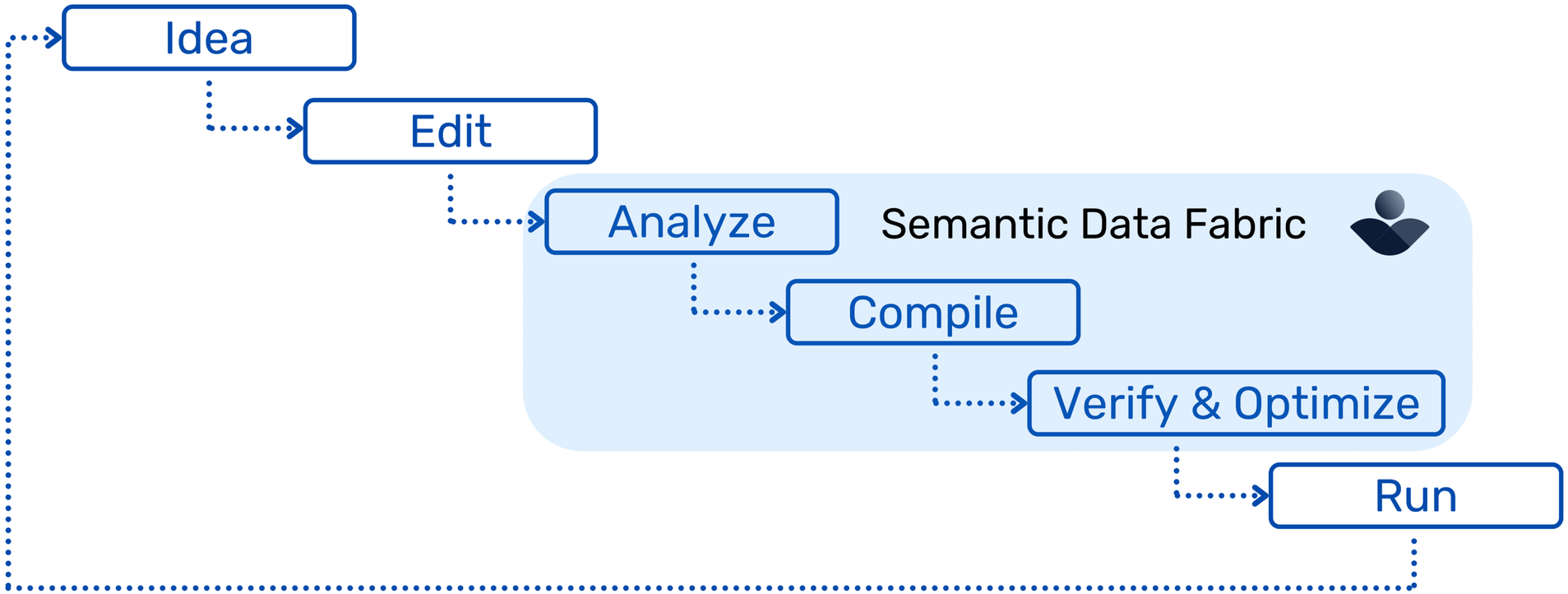
# SDF & Datafusion

Lukas Schulte (lukas@sdf.com)
Bo Lin (bo@sdf.com)

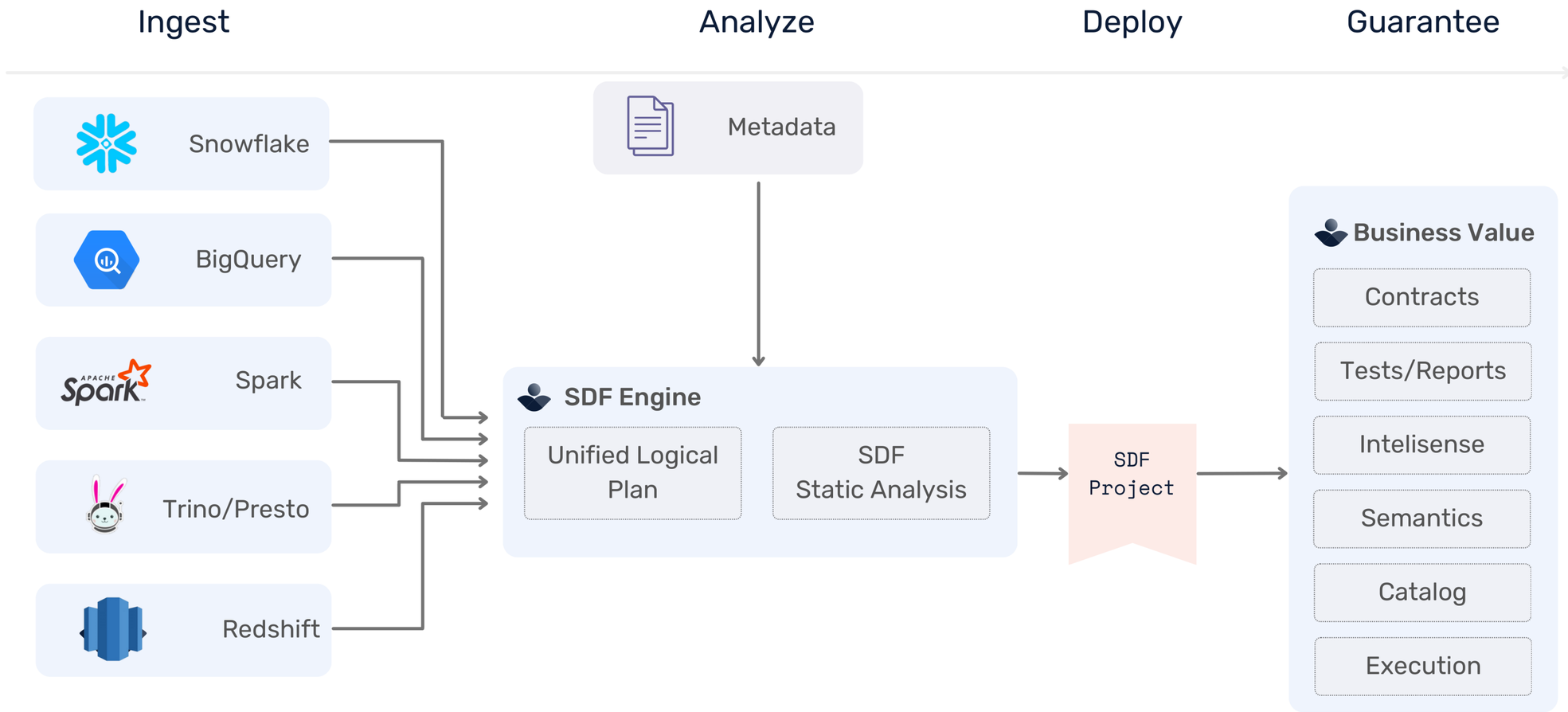
March 25, 2024

# What we do

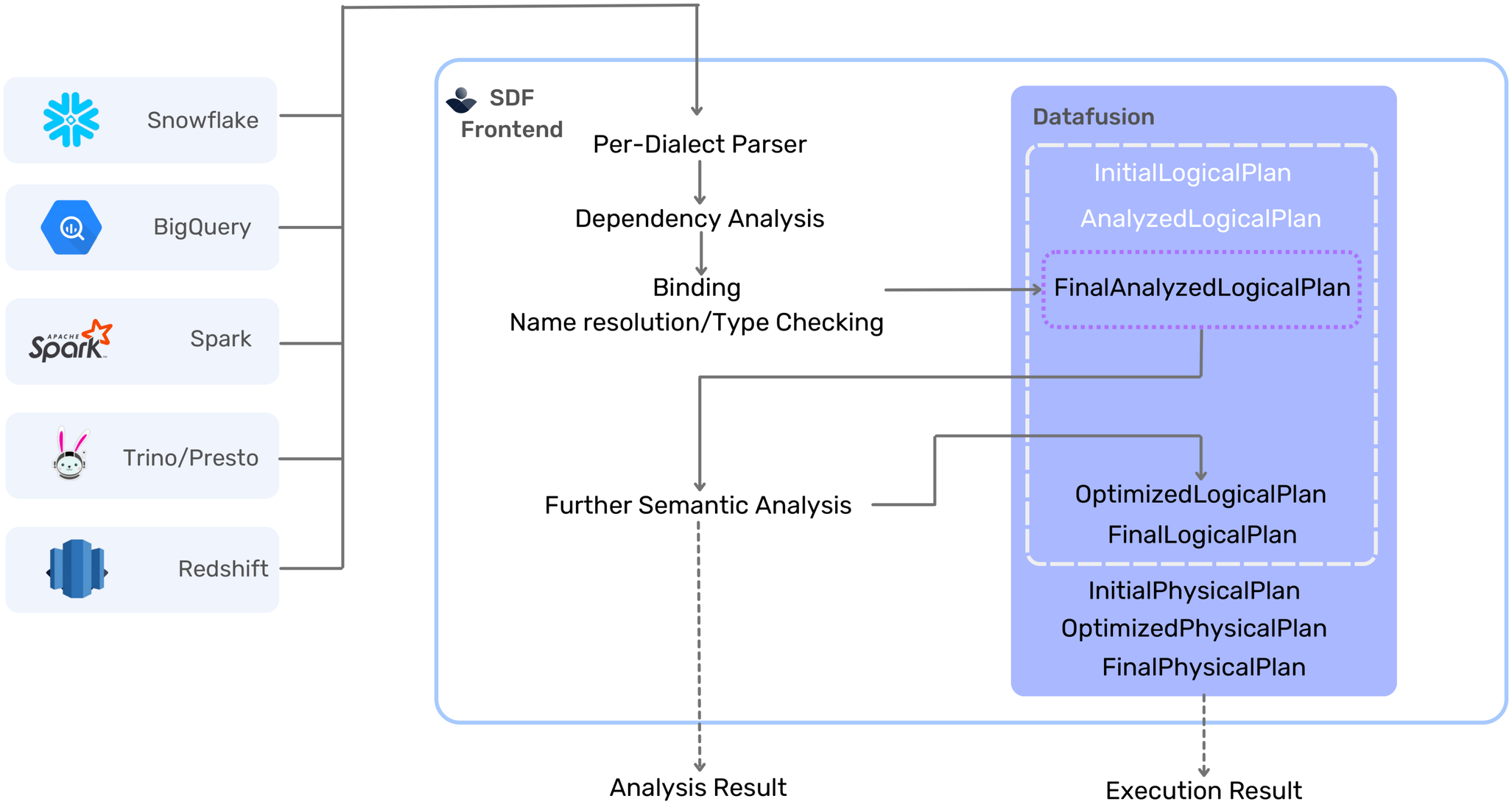
*Driving Fastest Cycle Time for Data Development*



# SDF Architecture



# How We Use Datafusion



# Improving Datafusion for SQL Analysis

**Performance**

Column resolution

Expr copy/hashing

**Usability**

`LogicalPlan` as “dumb structs”

**Stability**

Clear Separation of Frontend to Backend

Specialize mid-level IR

**What is the role of Substrait?**

# Other Areas We Hope to Contribute

## Functions

Documentation & Implementation

Dialect Specific Functions Crates With Code-Gen

- 738 Trino Functions
- 555 BigQuery Functions
- 268 Redshift Functions
- 1321 Snowflake Functions

## LogicalPlan to SQL/Substrait Translation

## Python Bindings

Rust-Python Bindings

- Defining Higher-Order Functions
- Defining Table Functions

functions.sdf.yml

```yaml
function:
  name: all_match
  parameters:
  - datatype: array<$1>
  - datatype: function($1, boolean)
  optional-parameters: []
  returns:
    datatype: boolean

---
function:
  name: approx_most_frequent
  kind: aggregate
  parameters:
  - datatype: bigint
  - datatype: varchar
  - datatype: bigint
  optional-parameters: []
  returns:
    datatype: map<varchar, bigint>
```

# Thank You

Lukas Schulte (lukas@sdf.com)
Bo Lin (bo@sdf.com)