*Data and Database Management in Algorithmic Trading Platform*

# Hedge Fund Data Sources: Paid vs. Free

*Comprehensive breakdown of data types and sources used by hedge funds*

## **1. Market & Trading Data**

### **Paid Sources**

* Bloomberg Terminal (Bloomberg Professional) – Real-time pricing, news, analytics ($24k+/year)
* Refinitiv Eikon / LSEG Workspace – Market data, research, and analytics (~$20k+/year)
* FactSet – Portfolio analytics, risk metrics, and alternative data integration (~$12k+/year)
* S&P Global Market Intelligence – Fundamental and credit data
* ICE Data Services – Fixed income, derivatives, and reference data
* Nasdaq TotalView – Ultra-low-latency Level 2 order book data (for HFT firms)

### **Free Sources**

* Yahoo Finance API – Free EOD (end-of-day) stock prices
* Alpha Vantage – Free (limited) real-time and historical market data
* Quandl (now part of Nasdaq) – Some free datasets (e.g., futures, commodities)
* FRED (Federal Reserve Economic Data) – Macroeconomic indicators

## **2. Alternative Data (Non-Traditional Datasets)**

### **Paid Sources**

* Satellite & Geolocation Data  
  + *Orbital Insight* – Retail foot traffic, oil tanker tracking
  + *RS Metrics* – Parking lot occupancy for retail predictions
* Credit Card & Consumer Spending  
  + *Second Measure (Yodlee)* – Aggregated credit card transactions
  + *Earnest Research* – Consumer spending trends
* Web Scraping & Sentiment  
  + *Thinknum Alternative Data* – Scrapes job postings, app downloads
  + *Dataminr* – Real-time event detection from social media
* Supply Chain & Shipping  
  + *Panjiva (S&P Global)* – Import/export shipping manifests
  + *MarineTraffic* – Real-time vessel tracking

### **Free (or Partially Free) Sources**

* Google Trends – Search volume trends for stocks/products
* Twitter/X API – Social sentiment analysis (limited free tier)
* SEC Edgar Database – Free filings (10-K, 10-Q, insider trades)
* UN Comtrade Database – Free global trade data

## **3. Fundamental & Company Data**

### **Paid Sources**

* Capital IQ (S&P Global) – Deep financial statement data
* Morningstar Direct – Mutual fund and ETF holdings
* MSCI ESG Research – Sustainability metrics
* Sentieo (ex-AlphaSense) – Earnings call transcripts, keyword search

### **Free Sources**

* SEC.gov (Edgar) – Free company filings
* Investopedia – Basic financial metrics (limited)
* MacroTrends – Free historical financials

## **4. News & Sentiment Data**

### **Paid Sources**

* RavenPack – AI-processed news sentiment scoring
* Bloomberg News / Reuters News Feed – Real-time financial news
* Dow Jones Newswires – High-impact news alerts

### **Free Sources**

* NewsAPI – Free (limited) news aggregation
* Reddit (WallStreetBets, etc.) – Crowdsourced sentiment (via web scraping)

## **5. Risk & Portfolio Analytics Data**

### **Paid Sources**

* Barra / MSCI Risk Models – Portfolio risk factor analysis
* Axioma – Multi-asset risk analytics
* BlackRock Aladdin – Institutional risk management

### **Free Sources**

* Portfolio Visualizer – Basic backtesting (free tier)
* QuantConnect – Free (limited) quant research tools

## **6. Macroeconomic & Geopolitical Data**

### **Paid Sources**

* Haver Analytics – Deep macroeconomic datasets
* Oxford Economics – Global forecasts
* IHS Markit (now S&P Global) – Economic indicators

### **Free Sources**

* World Bank Open Data
* IMF Data
* Trading Economics (free tier)

### **Summary Table: Hedge Fund Data Sources**

| Data Type | Paid Sources | Free Sources |
| --- | --- | --- |
| Market Data | Bloomberg, Refinitiv, FactSet | Yahoo Finance, Alpha Vantage |
| Alternative Data | Orbital Insight, Second Measure | Google Trends, SEC Edgar |
| Fundamental Data | Capital IQ, Morningstar | SEC.gov, MacroTrends |
| News/Sentiment | RavenPack, Dow Jones | NewsAPI, Reddit |
| Risk Analytics | Barra, Axioma, Aladdin | Portfolio Visualizer |
| Macro Data | Haver Analytics, Oxford Economics | World Bank, IMF |

### **Key Takeaways**

* Quant funds → Heavy reliance on *paid alternative data* (satellite, credit card, web scraping).
* Fundamental investors → *Bloomberg, Capital IQ, SEC filings*.
* Macro hedge funds → *Geopolitical/economic data* (Haver, Oxford Economics).

**1. Planning**

| **Task** | **Details / Examples** | **Priority** | **Status** |
| --- | --- | --- | --- |
| Define Trading Strategy | Momentum, mean reversion, arbitrage, etc. | High | ⬜ |
| Identify Data Types | OHLCV, tick, fundamentals, sentiment, alternative data | High | ⬜ |
| Choose DB Type | PostgreSQL, MongoDB, InfluxDB, TimescaleDB | High | ⬜ |
| Design Data Schema | Tables: symbols, candles, trades, features, signals | High | ⬜ |
| Define Data Granularity | Tick, 1-min, hourly, daily | Medium | ⬜ |

**2. Data Source**

| **Category** | **Task** | **Description** | **Examples / Tools Used** | **Priority** | **Completed** |
| --- | --- | --- | --- | --- | --- |
| Market Data | Select Market Data Providers | Choose sources for market data | Binance, Yahoo Finance, Alpha Vantage, Polygon.io | High | ⬜ |
| Fundamentals | Choose Fundamental Data Sources | Choose sources for company fundamentals | Yahoo, EDGAR, Quandl, IEX Cloud | Medium | ⬜ |
| News & Sentiment | Integrate News/Sentiment Sources | Connect to APIs for sentiment/news analysis | Twitter API, GDELT, NewsAPI, RavenPack | Medium | ⬜ |
| Historical Data | Download Historical Data | Use Python tools to gather historical data | yfinance, ccxt, ib\_insync | High | ⬜ |
| Data Storage | Store Raw Data Temporarily | Store raw data in flexible file formats before ingestion | CSV, Parquet, JSON | Medium | ⬜ |

**3. Database Setup**

| **Task** | **Description** | **Examples / Tools Used** | **Priority** | **Completed** |
| --- | --- | --- | --- | --- |
| Create DB Instance | Set up a local or cloud-hosted database | AWS RDS, MongoDB Atlas, etc. | High | ⬜ |
| Build Schema & Indexes | Normalize tables and apply efficient indexing | Time indexes, relational schema normalization | High | ⬜ |
| Optimize for Time-Series Data | Improve performance for time-series workloads | Partitioning, TimescaleDB hypertables | Medium | ⬜ |

**4. ETL Pipeline**

| **Task** | **Description** | **Examples / Tools Used** | **Priority** | **Completed** |
| --- | --- | --- | --- | --- |
| Build ETL Scripts | Extract, clean, resample, load | Python, Pandas, NumPy | High | ⬜ |
| Use Scheduler | Automate workflows | Cron jobs, Airflow, Prefect | High | ⬜ |
| Validate Data | Check data quality and integrity | NA handling, outlier checks, logging | High | ⬜ |
| Store Processed Data | Organize clean data for analysis/modeling | OHLCV tables, indicators, model-ready features | High | ⬜ |

**5. Backtesting Preparation**

| **Task** | **Description** | **Examples / Tools Used** | **Priority** | **Completed** |
| --- | --- | --- | --- | --- |
| Enable Data Exports | Pull historical samples per strategy | Export from DB or CSV for backtest engines | High | ⬜ |
| Feature Engineering Storage | Store technical indicators for models/backtests | SMA, EMA, RSI, Bollinger Bands, MACD, etc. | High | ⬜ |
| Simulated Trades Table | Save backtest results including performance | Orders, P&L, drawdowns | Medium | ⬜ |

**6. Realtime Data**

| **Task** | **Description** | **Examples / Tools Used** | **Priority** | **Completed** |
| --- | --- | --- | --- | --- |
| Stream Live Prices | Stream market data in real time | WebSockets: Binance, Alpaca, Interactive Brokers (IB) | High | ⬜ |
| Store Tick Data | Store tick-level data efficiently | Time and symbol indexing, optimized schema | Medium | ⬜ |
| Queue for Ingestion | Use message queues for scalable data ingestion | Kafka, Redis, RabbitMQ | Medium | ⬜ |

**7. Monitoring**

| **Task** | **Description** | **Examples / Tools Used** | **Priority** | **Completed** |
| --- | --- | --- | --- | --- |
| Monitor DB Performance | Track and visualize database health | PgAdmin, Grafana, Prometheus | Medium | ⬜ |
| Log Ingestion Jobs | Capture logs for ETL and ingestion processes | Python logging, DB logs, notifications | High | ⬜ |
| Alert on Failures | Set up real-time alerts for critical issues | Email, Slack, PagerDuty | Medium | ⬜ |

**8. Backup and Storage**

| **Task** | **Description** | **Examples / Tools Used** | **Priority** | **Completed** |
| --- | --- | --- | --- | --- |
| Schedule Backups | Perform regular backups | Daily dumps, cloud backups | High | ⬜ |
| Store on Cloud | Use cloud storage for backup/data archiving | AWS S3, GCS, Azure Blob | Medium | ⬜ |
| Apply Retention Policies | Manage data lifecycle | Delete raw data after 90 days | Medium | ⬜ |

**9. Security**

| **Task** | **Description** | **Examples / Tools Used** | **Priority** | **Completed** |
| --- | --- | --- | --- | --- |
| Manage API Keys Securely | Use secure methods to store API keys | .env, vaults, encryption | High | ⬜ |
| Use Role-Based DB Access | Implement different access levels in DB | Admin, read-only, analyst roles | High | ⬜ |
| Enable Encryption | Secure data in transit and at rest | In-transit (SSL), at-rest (AES-256) | High | ⬜ |