**Table of Contents**

1. Project Overview
   * Vision and Goals
   * Core Features and Benefits
   * Technical Architecture Overview
   * Project Scope and Limitations
2. System Architecture
   * Backend Infrastructure (Oracle Cloud)
   * Frontend Infrastructure (Vercel)
   * Data Flow and Integration Points
   * System Components Interaction
3. Data Pipeline and Processing
   * Data Collection Strategy
   * ETL Pipeline Design with Dagster
   * Data Storage Architecture
   * News Processing and Analysis Pipeline
4. Database Design and Implementation
   * PostgreSQL Schema Design
   * Neo4j Graph Database Structure
   * Data Relationships and Connections
   * Query Optimization Strategies
5. Frontend Implementation
   * Vue.js Application Structure
   * Vercel Serverless Functions Design
   * API Layer Architecture
   * User Interface Components
6. Features Implementation Details
   * TL;DR Generation System
   * Knowledge Graph Visualization
   * Social Media Integration
   * Video Content Analysis
   * Search and Navigation
7. Deployment and Operations
   * Oracle Cloud Setup and Configuration
   * Vercel Deployment Process
   * Monitoring and Logging
   * Backup and Recovery Strategies

**1. Project Overview**

**Vision and Goals**

This project aims to revolutionize news consumption by creating a comprehensive news aggregation and analysis platform that brings together content from multiple sources, processes it intelligently, and presents it in an easily digestible format. The platform will focus on English-language content from the top 20 most visited news portals worldwide.

Key Goals:

* Eliminate information overload by providing concise, fact-based summaries
* Reveal connections between news stories through knowledge graphs
* Provide multi-dimensional analysis including social media sentiment and video content
* Ensure scalability and reliability through modern cloud infrastructure

**Core Features and Benefits**

**1. News Aggregation and Analysis**

Features:

* Automated collection from top 20 global news portals
* Organization by portal and category (politics, sports, tech, etc.)
* Cross-source fact verification and comparison
* Extraction of key entities, actions, and relationships

Benefits:

* Comprehensive coverage from reliable sources
* Reduced time spent on news consumption
* Enhanced understanding through cross-source verification

**2. TL;DR System**

Features:

* Twitter-like concise summaries
* Most important words found in core 5 news sites
* Key facts extraction from each source
* Differences highlighting across sources

Benefits:

* Quick understanding of complex stories
* Easy comparison of different news sources
* Time-saving for busy readers

**3. Knowledge Graph Visualization**

Features:

* Visual representation of news relationships
* Entity linking (subjects, actions, organizations)
* Interactive exploration of connections
* Temporal relationship tracking

Benefits:

* Understanding complex news relationships
* Discovering hidden connections
* Better context for news stories

**4. Social Media Integration**

Features:

* Aggregated sentiment analysis from top comments
* Keyword extraction from social discussions
* Top 5 most liked comments per story
* Trend analysis across platforms

Benefits:

* Understanding public reaction
* Identifying key discussion points
* Tracking story evolution on social media

**5. Video Content Analysis**

Features:

* Related video aggregation
* YouTube comment analysis
* Sentiment analysis of video reactions
* Key moment identification

Benefits:

* Multi-media perspective on news
* Understanding video content context
* Access to public video reactions

**Technical Architecture Overview**

**Backend Infrastructure (Oracle Cloud)**

* Oracle Linux 8
* VM.Standard.A2.Flex (2 OCPU, 12GB RAM)
* Block storage for data persistence
* Hosting:
  + Dagster ETL pipelines
  + PostgreSQL database
  + Neo4j graph database

**Frontend Infrastructure (Vercel)**

* Vue.js frontend application
* Vercel Serverless Functions for API
* Automated deployment and scaling
* Global CDN distribution

**Project Scope and Limitations**

**In Scope**

* English language news sources only
* Top 20 most visited news portals
* Major news categories (politics, sports, tech, etc.)
* Public social media content
* Publicly available video content

**Out of Scope**

* Non-English content
* Real-time news updates
* Private or subscription-based content
* Historical news archive (beyond defined retention period)
* User-generated content moderation

**Vision and Goals - Detailed Breakdown**

**Primary Vision**

To create a revolutionary news consumption platform that transforms how users interact with and understand global news by:

* Eliminating information overload through intelligent aggregation and summarization
* Providing deeper insights through multi-dimensional analysis
* Creating visual connections between related news items
* Offering fact-based, unbiased news understanding

**Strategic Goals**

**1. Information Aggregation Excellence**

* **Comprehensive Coverage**:
  + Integration with top 20 English-language news portals
  + Systematic categorization across major news categories
  + Unified access to dispersed news content
* **Smart Organization**:
  + Hierarchical category structure for easy navigation
  + Cross-referencing between related stories
  + Temporal organization of news evolution

**2. Content Processing and Analysis**

* **Automated Processing Pipeline**:
  + Real-time scraping and processing of news articles
  + Intelligent categorization and tagging
  + Entity and relationship extraction
* **Fact Extraction and Verification**:
  + Cross-source fact validation
  + Identification of common narratives
  + Highlighting of divergent viewpoints

**3. User Experience Enhancement**

* **Efficient Information Consumption**:
  + Quick access to essential information through TL;DR summaries
  + Visual representation of complex relationships
  + Intuitive navigation between related content
* **Multi-dimensional Understanding**:
  + Integration of social media perspectives
  + Video content context
  + Public sentiment analysis

**4. Technical Excellence**

* **Scalable Architecture**:
  + Ability to handle growing number of sources
  + Efficient processing of increasing data volumes
  + Flexible infrastructure for feature expansion
* **Reliable Performance**:
  + Consistent response times
  + High availability
  + Robust error handling

**5. Data Intelligence**

* **Smart Analysis**:
  + Pattern recognition in news coverage
  + Trend identification
  + Relationship mapping between stories
* **Knowledge Graph Building**:
  + Entity relationship mapping
  + Temporal relationship tracking
  + Interactive exploration capabilities

**Measurable Objectives**

**Performance Metrics**

* Processing time for new articles < 5 minutes
* TL;DR generation accuracy > 95%
* System uptime > 99.9%
* API response time < 200ms

**User Experience Metrics**

* Time to access key information < 30 seconds
* Navigation depth to related content ≤ 3 clicks
* Complete story understanding time reduced by 70%

**Data Quality Metrics**

* Cross-source fact verification accuracy > 95%
* Entity extraction precision > 90%
* Relationship mapping accuracy > 85%

**Long-term Vision Elements**

**1. Platform Evolution**

* Expansion to additional languages
* Integration with emerging news sources
* Advanced AI-driven analysis capabilities

**2. Community Building**

* Collaborative fact-checking
* Expert insights integration
* Community-driven knowledge enhancement

**3. Technology Innovation**

* Advanced natural language processing
* Enhanced visualization capabilities
* Improved relationship mapping

This vision drives towards creating a platform that not only aggregates news but transforms it into actionable, understandable knowledge for users while maintaining high standards of accuracy and reliability.

**News Aggregation and Analysis - Detailed Breakdown**

**Collection and Storage Architecture**

**Raw Data Collection**

* **Source Management**
  + Top 20 English-language news portals
  + Organized in Oracle Cloud Storage buckets
  + Structure: portal\_name/category/raw\_data
  + Example portals: BBC, CNN, Reuters, NYT, etc.

**Category Organization**

* **Primary Categories**
  + Politics (National/International)
  + Economics & Business
  + Technology & Science
  + Sports
  + Entertainment
  + Health & Wellness
  + Environmental News
  + Education
  + World News
  + Cultural Affairs

**Analysis Components**

**1. Content Extraction and Processing**

* **Text Extraction**
  + Removal of advertisements and irrelevant content
  + Extraction of main article body
  + Preservation of critical metadata (author, publication date, etc.)
* **Structural Analysis**
  + Identification of article components (headlines, subheadings, quotes)
  + Recognition of multimedia elements
  + Extraction of embedded references and links

**2. Cross-Source Verification**

* **Fact Comparison**
  + Identification of common facts across sources
  + Detection of contradictions or discrepancies
  + Verification through multiple source confirmation
* **Source Credibility Assessment**
  + Historical accuracy tracking
  + Citation and reference verification
  + Expert source identification

**3. Entity Extraction**

* **Named Entity Recognition**
  + People identification
  + Organization detection
  + Location extraction
  + Date and time recognition
* **Relationship Mapping**
  + Entity connections
  + Action attribution
  + Temporal relationships
  + Causal relationships

**4. Content Classification**

* **Topic Classification**
  + Multi-label categorization
  + Sub-topic identification
  + Theme extraction
* **Sentiment Analysis**
  + Overall tone assessment
  + Bias detection
  + Emotional content evaluation

**Benefits Implementation**

**1. Time Efficiency**

* **Quick Access**
  + Instant access to verified facts
  + Reduced research time
  + Efficient cross-source comparison
* **Information Density**
  + High-value content prioritization
  + Elimination of redundant information
  + Focus on unique insights

**2. Quality Assurance**

* **Accuracy Enhancement**
  + Multi-source verification
  + Fact-checking automation
  + Contradiction highlighting
* **Bias Mitigation**
  + Multiple perspective presentation
  + Balanced viewpoint representation
  + Source diversity tracking

**3. User Empowerment**

* **Customization Options**
  + Personal interest focus
  + Category-based filtering
  + Source preference settings
* **Interactive Exploration**
  + Related content discovery
  + Deep-dive capabilities
  + Historical context access

**Technical Implementation**

**ETL Pipeline (Dagster)**

1. **Data Collection Jobs**
   * Scheduled scraping tasks
   * Source monitoring
   * Update detection
2. **Processing Jobs**
   * Content extraction
   * Entity recognition
   * Relationship mapping
3. **Analysis Jobs**
   * Cross-source verification
   * Sentiment analysis
   * Classification tasks

**Storage Architecture**

1. **Raw Data Storage**
   * Oracle Cloud Storage buckets
   * Organized by source and category
   * Version control for updates
2. **Processed Data Storage**
   * PostgreSQL for structured data
   * Neo4j for relationship graphs
   * Cache layer for frequent access

**API Integration**

1. **Data Access Endpoints**
   * Content retrieval
   * Search functionality
   * Filter operations
2. **Analysis Endpoints**
   * Entity information
   * Relationship queries
   * Verification status

**TL;DR System - Detailed Breakdown**

**Core Purpose and System Overview**

The TL;DR (Too Long; Didn't Read) system serves as an intelligent summarization engine that processes news articles from multiple sources to create concise, fact-focused summaries. Unlike traditional summarization tools, this system specifically focuses on cross-source comparison and fact extraction, ensuring that users receive the most crucial information while highlighting different perspectives from various news outlets.

**Content Processing Mechanism**

**Initial Content Analysis**

The system begins by analyzing articles covering the same news story from the top 5 news sources. Each article undergoes deep linguistic analysis to identify key components such as main events, primary actors, critical quotes, and essential facts. During this phase, the system employs Natural Language Processing (NLP) to understand the context and importance of each piece of information.

**Key Word Identification and Weighting**

The system implements a sophisticated word importance algorithm that goes beyond simple frequency counting. It identifies words and phrases that appear prominently across multiple sources, with special attention to:

The placement of words within articles (headlines, first paragraphs, conclusions) The context in which words appear (quotes from key figures, statistical data, event descriptions) The relationship between words and identified entities (people, organizations, locations)

This weighting mechanism helps determine which elements are truly crucial to the story versus what might be editorial or supplementary content.

**Summary Generation Process**

**Cross-Source Fact Extraction**

The system analyzes how facts are presented across different sources, identifying:

Common narrative elements that appear consistently across sources Variations in how different sources present the same information Unique perspectives or additional context provided by specific sources

This cross-referencing helps establish a reliability score for each piece of information and ensures that the final summary includes verified facts rather than speculation or opinion.

**Twitter-Style Summary Creation**

The system generates concise, Twitter-like summaries that capture the essence of the news story. These summaries are structured to:

Lead with the most impactful and verified information Include numerical data or statistics when relevant Highlight any significant disagreements between sources Maintain readability while maximizing information density

**Differential Analysis**

A key feature of the system is its ability to identify and highlight differences in reporting across sources. This includes:

Variations in emphasis placed on different aspects of the story Differences in factual claims or numerical data Contrasting interpretations of events or statements

**Integration with Knowledge Graph**

The TL;DR system doesn't operate in isolation but is deeply integrated with the platform's knowledge graph. This integration allows for:

Contextual enrichment of summaries with related historical events Identification of connections to ongoing news stories Enhancement of summaries with relevant entity relationships

**User Interface and Presentation**

The presentation layer of the TL;DR system is designed to provide maximum value with minimal cognitive load. Users can:

View the base summary for immediate understanding Expand to see source-specific variations Access the full articles for deeper reading Explore related summaries through knowledge graph connections

**Quality Assurance and Verification**

The system maintains high accuracy through:

Continuous cross-referencing of facts across sources Verification against trusted reference sources Regular updating as new information becomes available User feedback integration for improvement

The TL;DR system essentially acts as a smart news curator, providing users with verified, concise information while maintaining the complexity and nuance of the original reporting. It saves users significant time while ensuring they don't miss crucial details or important perspectives on the news.

**Knowledge Graph Visualization - Detailed Breakdown**

**Core Purpose and Architecture**

The Knowledge Graph Visualization system transforms complex news data into an intuitive, interactive visual representation of interconnected information. It serves as a dynamic map of news stories, entities, and their relationships, enabling users to understand complex news landscapes at a glance while providing tools for deep exploration of connections and patterns.

**Graph Structure and Components**

**Entity Representation**

The system identifies and represents various types of entities within the news ecosystem:

Primary Entities:

* News Articles: Core nodes representing individual news stories
* People: Key figures mentioned across articles
* Organizations: Companies, governments, institutions
* Locations: Geographical contexts of stories
* Events: Significant occurrences or happenings
* Topics: Broader themes or subjects

These entities are stored in Neo4j with rich metadata including temporal information, source verification status, and relevance scores.

**Relationship Mapping**

The system maintains sophisticated relationship types between entities:

Direct Relationships: The system captures explicit connections mentioned in news articles, such as "Person X leads Organization Y" or "Organization A acquires Organization B"

Inferred Relationships: Using advanced NLP and pattern recognition, the system identifies implicit connections between entities even when not directly stated in articles

Temporal Relationships: All connections are timestamped and versioned, allowing users to see how relationships evolve over time

**Visual Representation System**

**Dynamic Layout Engine**

The visualization engine employs sophisticated layout algorithms that:

* Automatically organize nodes based on relationship strength and relevance
* Adjust layout density based on viewing context and zoom level
* Highlight important nodes and relationships based on user context
* Maintain visual clarity even with complex relationship networks

**Interactive Features**

Users can interact with the graph through:

Navigation Controls:

* Zoom levels for different detail granularity
* Pan and scroll across the knowledge landscape
* Focus+Context viewing for maintaining orientation

Selection and Filtering:

* Entity type filtering
* Temporal range selection
* Relationship type filtering
* Source-based filtering

**Intelligence Layer**

**Pattern Recognition**

The system employs advanced analytics to:

Identify significant patterns in news coverage and relationships Detect emerging trends and connections Highlight unusual or noteworthy relationship patterns Track the evolution of stories and relationships over time

**Contextual Enhancement**

The visualization is enriched with:

Historical Context:

* Previous related events
* Historical patterns
* Past relationships between entities

Current Context:

* Real-time updates
* Breaking news integration
* Social media sentiment overlay

**Technical Implementation**

**Backend Processing**

The system utilizes:

Neo4j for graph storage and query processing:

* Optimized graph algorithms for relationship analysis
* Efficient query patterns for real-time visualization
* Scalable storage architecture for growing data

Custom Processing Engine:

* Entity relationship extraction
* Pattern recognition algorithms
* Real-time update processing

**Frontend Rendering**

The visualization layer employs:

Modern WebGL-based rendering:

* Efficient handling of large graph structures
* Smooth animations and transitions
* Responsive interaction handling

Adaptive Display:

* Mobile-friendly layouts
* Resolution-independent rendering
* Accessibility features

**Integration Points**

**Data Pipeline Integration**

The visualization system integrates with:

* The TL;DR system for summary context
* Social media analysis for sentiment overlay
* Video content analysis for multimedia context
* Real-time news updates for current information

**API Layer**

The system exposes:

* Graph query endpoints for custom visualizations
* Real-time update webhooks
* Custom filter endpoints
* Export capabilities for external use

This comprehensive knowledge graph visualization system transforms complex news relationships into an intuitive, interactive experience, enabling users to understand and explore news connections in ways previously impossible with traditional news consumption methods.

**Social Media Integration - Detailed Breakdown**

**Core Purpose and System Overview**

The Social Media Integration system aggregates, analyzes, and presents social media reactions and discussions related to news stories. It provides a comprehensive view of public opinion, sentiment trends, and key discussion points across major social media platforms, transforming unstructured social conversations into actionable insights.

**Data Collection and Processing Framework**

**Platform Coverage and Integration**

The system focuses on major social platforms that generate significant news-related discussions:

Primary Sources:

* Twitter: Real-time reactions and trending discussions
* Reddit: In-depth community discussions and analysis
* LinkedIn: Professional perspectives and industry insights
* YouTube Comments: Video-related discussions
* Facebook Public Posts: Wide-reaching public discussions

**Collection Methodology**

The system implements sophisticated data collection strategies:

Real-time Monitoring: The system continuously monitors social media platforms for relevant discussions, using intelligent filtering to identify content related to specific news stories. This includes tracking hashtags, keywords, and engagement patterns.

Content Categorization: Collected content is automatically categorized based on:

* Relevance to the news story
* Platform of origin
* Content type (comment, share, original post)
* User engagement metrics
* Timestamp and context

**Analysis Components**

**Sentiment Analysis Engine**

The system performs multi-layered sentiment analysis:

Deep Sentiment Processing:

* Beyond basic positive/negative classification
* Detection of nuanced emotional responses
* Recognition of sarcasm and implicit sentiment
* Context-aware sentiment scoring

Aggregate Sentiment Metrics:

* Overall sentiment trends
* Platform-specific sentiment patterns
* Temporal sentiment evolution
* Demographic-based sentiment analysis

**Top Comments Selection**

The system identifies and presents the most significant comments using a sophisticated ranking algorithm that considers:

Engagement Metrics:

* Like counts and reactions
* Reply threads and discussion quality
* Share counts and reach
* User credibility and history

Content Quality:

* Informativeness and relevance
* Unique insights or perspectives
* Supporting evidence or links
* Writing quality and clarity

**Integration with News Analysis**

**Cross-Platform Analysis**

The system correlates social media activity with news coverage:

Temporal Correlation:

* Tracking how social media discussion evolves with news developments
* Identifying peaks in social media activity
* Mapping discussion trends to news events

Content Correlation:

* Matching social media topics with news themes
* Identifying new angles emerging from social discussions
* Tracking information spread across platforms

**Keyword and Topic Extraction**

Advanced natural language processing identifies:

Emerging Topics:

* New discussion themes
* Trending hashtags
* Evolving narratives

Key Discussion Points:

* Common arguments and viewpoints
* Frequently cited sources
* Contested information

**User Interface and Presentation**

**Interactive Dashboard**

The system presents social media insights through:

Visual Analytics:

* Sentiment trend graphs
* Engagement heat maps
* Topic evolution timelines
* Platform comparison charts

Content Browsers:

* Curated top comments view
* Thread exploration tools
* Cross-platform discussion tracking
* Real-time update feeds

**Integration Features**

The dashboard integrates with other system components:

Knowledge Graph Connection:

* Social media nodes in the knowledge graph
* Relationship visualization with news entities
* User exploration pathways

TL;DR Enhancement:

* Social context for news summaries
* Public reaction highlights
* Key discussion points integration

**Privacy and Ethical Considerations**

**Data Handling**

The system implements strict privacy controls:

Content Filtering:

* Personal information removal
* Sensitive content screening
* Age-appropriate content filtering

Attribution Management:

* Proper source attribution
* User privacy protection
* Platform compliance

This comprehensive social media integration system transforms scattered social discussions into structured, meaningful insights that enhance understanding of news impact and public reception.

**Video Content Analysis - Detailed Breakdown**

**Core Purpose and System Overview**

The Video Content Analysis system discovers, processes, and integrates video content related to news stories across major platforms, primarily focusing on YouTube. It enriches news understanding by providing visual context, analyzing video comments and reactions, and extracting key moments and insights from video content.

**Video Content Discovery and Integration**

**Content Sourcing**

The system actively identifies relevant video content through:

Source Prioritization:

* Official news channel videos
* Verified content creator coverage
* Expert analysis videos
* On-the-ground footage
* Press conferences and official statements

Discovery Mechanisms:

* Keyword and topic matching
* Channel monitoring for major news outlets
* Trending video detection
* Cross-reference with news articles
* Timestamp correlation with news events

**Analysis Components**

**Video Content Processing**

The system analyzes video content across multiple dimensions:

Metadata Analysis:

* Upload timing relative to news events
* Channel credibility assessment
* View counts and engagement metrics
* Video description and tag analysis
* Related video connections

Content Analysis:

* Key moment identification
* Speech-to-text transcription
* Visual element detection
* Caption and subtitle extraction
* Thumbnail analysis for relevance

**Comment Analysis System**

Deep Comment Processing:

* Sentiment classification of comments
* Key discussion point extraction
* Expert comment identification
* Timestamp-linked comments analysis
* Controversy detection

Engagement Metrics:

* Like/dislike ratios
* Comment engagement patterns
* Share statistics
* Viewer retention data
* Interaction timing analysis

**Integration Features**

**News Story Enhancement**

Videos are integrated with news stories through:

Contextual Linking:

* Relevant video segments linked to specific article points
* Timeline integration with news development
* Source comparison and verification
* Multiple perspective presentation

Content Enrichment:

* Visual evidence for news claims
* Expert commentary integration
* Event footage correlation
* Press statement integration

**Knowledge Graph Integration**

Video nodes in the knowledge graph include:

* Connections to related news articles
* Links to key figures and organizations
* Temporal relationship mapping
* Topic and theme connections
* Source credibility indicators

**User Interface Elements**

**Video Content Presentation**

Interactive Video Browser:

* Thumbnail preview grid
* Relevance-based sorting
* Platform source filtering
* Timeline-based navigation
* Key moment highlighting

Content Navigation:

* Jump to specific segments
* Side-by-side comparison
* Transcript-based navigation
* Related video suggestions

**Analysis Display**

Insight Dashboard:

* Comment sentiment trends
* Key discussion points
* Engagement metrics visualization
* Temporal analysis charts
* Source credibility indicators

**Technical Implementation**

**Video Processing Pipeline**

Content Processing:

* Automated video download and storage
* Transcript generation and processing
* Visual element detection
* Key frame extraction
* Metadata extraction and indexing

Comment Processing:

* Continuous comment collection
* Sentiment analysis processing
* Engagement metric tracking
* Discussion thread analysis
* Update monitoring

**Storage and Retrieval**

Data Management:

* Efficient video metadata storage
* Comment data indexing
* Analysis result caching
* Relationship mapping storage
* Update tracking system

**Quality Assurance**

Content Verification:

* Source authenticity checking
* Content accuracy verification
* Engagement pattern analysis
* Spam and bot detection
* Misleading content identification

**Performance Optimization**

Resource Management:

* Selective video processing
* Incremental update processing
* Cache optimization
* Storage efficiency
* Bandwidth optimization

This comprehensive video content analysis system transforms video content into structured, searchable, and meaningful components of the news understanding ecosystem, providing crucial visual context and additional perspectives to news stories.

**Technical Architecture Overview**

**Backend Infrastructure (Oracle Cloud) - Initial Development Configuration and Scaling Plan**

**Initial Development Setup (Single VM Configuration)**

**VM Specifications**

* **Operating System**: Oracle Linux 8
* **Shape**: VM.Standard.A2.Flex
  + 2 OCPU cores
  + 12GB RAM
  + 2 Gbps Network Bandwidth
  + Block Storage Only

**Storage Configuration**

1. **Boot Volume**:
   * 50 GB (Default)
   * Oracle Linux 8 system files
2. **Block Storage Volumes**:
   * Data Volume (200GB):
     + PostgreSQL database files
     + Neo4j database files
     + Temporary processing files
   * Application Volume (100GB):
     + Application code
     + Dagster pipelines
     + Logs and monitoring data

**Component Distribution**

All components initially running on single VM:

1. **Dagster ETL Pipeline**
   * Resource Allocation:
     + 0.5 OCPU
     + 4GB RAM
   * Components:
     + Dagster Daemon
     + Dagster Webserver
     + Pipeline workers
2. **PostgreSQL Database**
   * Resource Allocation:
     + 0.5 OCPU
     + 4GB RAM
   * Configuration:
     + max\_connections: 100
     + shared\_buffers: 2GB
     + effective\_cache\_size: 3GB
3. **Neo4j Graph Database**
   * Resource Allocation:
     + 0.5 OCPU
     + 3GB RAM
   * Configuration:
     + dbms.memory.heap.initial\_size: 2G
     + dbms.memory.heap.max\_size: 2G
     + dbms.memory.pagecache.size: 1G
4. **System Resources**
   * Resource Allocation:
     + 0.5 OCPU
     + 1GB RAM reserved for OS operations

**Scaling Plan**

**Phase 1: Initial Development (Current)**

* Single VM setup as described above
* Focus on feature development and basic functionality
* Limited data processing capacity
* Estimated monthly cost: ~$100-150

**Phase 2: Component Separation (Month 3-4)**

When reaching processing bottlenecks or ~1000 active users:

1. **Database VM** (VM.Standard.A2.Flex)
   * 2 OCPU, 12GB RAM
   * PostgreSQL + Neo4j
   * Estimated cost: $150/month
2. **Application VM** (VM.Standard.A2.Flex)
   * 2 OCPU, 12GB RAM
   * Dagster + Application logic
   * Estimated cost: $150/month

**Phase 3: Database Separation (Month 6-7)**

When database load increases or ~5000 active users:

1. **PostgreSQL VM** (VM.Standard.A2.Flex)
   * 2 OCPU, 12GB RAM
   * Dedicated to PostgreSQL
   * Estimated cost: $150/month
2. **Neo4j VM** (VM.Standard.A2.Flex)
   * 2 OCPU, 12GB RAM
   * Dedicated to Neo4j
   * Estimated cost: $150/month
3. **Application VM** (Same as Phase 2)
   * Dagster + Application logic
   * Estimated cost: $150/month

**Phase 4: Processing Scale-Out (Month 9-10)**

When processing demands increase or ~10,000 active users:

1. **Database VMs** (Same as Phase 3)
   * PostgreSQL VM
   * Neo4j VM
2. **Application VM Cluster**
   * 2-3 VM.Standard.A2.Flex instances
   * Load balanced configuration
   * Estimated cost: $300-450/month
3. **Processing VM Cluster**
   * 2-3 VM.Standard.A2.Flex instances
   * Distributed Dagster workers
   * Estimated cost: $300-450/month

**Monitoring and Trigger Points**

**Performance Metrics**

* CPU Utilization > 70% sustained
* Memory Usage > 80% sustained
* Database Connection Pool > 70% utilized
* Processing Queue Backlog > 30 minutes
* Response Time > 500ms

**User Metrics**

* Active Users > 1000
* Concurrent Requests > 100
* Data Processing Volume > 1000 articles/day

**Cost Metrics**

* Monthly Infrastructure Cost
* Cost per User
* Processing Cost per Article

**Optimization Strategies**

1. **Resource Optimization**
   * Implementation of caching layers
   * Query optimization
   * Data archival strategies
   * Resource scheduling
2. **Cost Management**
   * Reserved instance pricing
   * Storage tiering
   * Resource auto-scaling
   * Performance monitoring

**Frontend Infrastructure (Vercel) - Development and Production Strategy**

**Development Architecture**

**Vue.js Application Structure**

Initial development setup organized in feature-based architecture:

Copy

src/

├── components/

│ ├── news/

│ │ ├── TldrSection.vue

│ │ ├── KnowledgeGraph.vue

│ │ ├── SocialFeed.vue

│ │ └── VideoAnalysis.vue

│ ├── shared/

│ │ ├── Loading.vue

│ │ └── ErrorBoundary.vue

│ └── layout/

├── composables/

│ ├── useNewsData.ts

│ ├── useGraphData.ts

│ └── useSocialData.ts

├── api/

│ ├── news.ts

│ ├── graph.ts

│ ├── social.ts

│ └── video.ts

└── pages/

├── index.vue

├── news/[id].vue

└── topic/[id].vue

**Vercel Serverless Functions (API Layer)**

Functions are co-located with the application code:

Copy

api/

├── news/

│ ├── getTldr.ts

│ ├── getArticle.ts

│ └── search.ts

├── graph/

│ ├── getRelations.ts

│ └── getEntities.ts

├── social/

│ ├── getSentiment.ts

│ └── getComments.ts

└── video/

├── getRelated.ts

└── getAnalysis.ts

**Development to Production Pipeline**

**Phase 1: Local Development**

1. **Development Environment**
   * Vue.js application with Vite
   * Local serverless function testing using Vercel CLI
   * Environment variables for different stages
2. **API Development**

typescript

Copy

*// api/news/getTldr.ts*

export default async function handler(req, res) {

const { articleId } = req.query

*// Connect to Oracle backend*

*// Process and return data*

}

**Phase 2: Feature Implementation**

1. **Core Components**
   * TL;DR View Implementation
   * Knowledge Graph Visualization
   * Social Media Feed
   * Video Analysis Component
2. **API Integration**
   * Each component gets its dedicated API endpoints
   * Serverless functions handle data fetching and processing
   * Caching implementation where appropriate

**Phase 3: Testing and Staging**

1. **Development Environment**
   * vercel dev for local testing
   * Points to development backend on Oracle Cloud
2. **Staging Environment**
   * Automatic deployments from staging branch
   * Points to staging backend
   * Preview deployments for PR reviews
3. **Production Environment**
   * Production deployments from main branch
   * Points to production backend
   * Zero-downtime deployments

**Vercel Serverless Functions Strategy**

**Data Fetching Layer**

typescript

Copy

*// api/news/getTldr.ts*

import { connectToNeo4j, connectToPostgres } from '../lib/db'

export default async function handler(req, res) {

try {

const data = await fetchNewsData(req.query)

const graph = await fetchGraphData(req.query)

return res.status(200).json({

tldr: data,

relations: graph

})

} catch (error) {

return res.status(500).json({ error: error.message })

}

}

**Caching Strategy**

* Implement caching at edge locations
* Use Vercel's Edge Cache
* Implement stale-while-revalidate pattern

**Error Handling**

* Consistent error response format
* Proper status codes
* Error tracking integration

**Scaling Considerations**

**Vercel's Built-in Scaling**

* Automatic scaling of serverless functions
* Global CDN distribution
* Edge caching and computation

**Performance Optimization**

1. **Image Optimization**
   * Use Vercel's Image Optimization
   * Automatic WebP conversion
   * Responsive images
2. **Code Splitting**
   * Route-based code splitting
   * Component lazy loading
   * Dynamic imports for heavy features
3. **Edge Functions**
   * Use edge functions for global performance
   * Implement region-based routing
   * Cache API responses at edge

**Monitoring and Analytics**

**Performance Monitoring**

* Vercel Analytics integration
* Custom performance metrics
* Real-time error tracking

**Usage Analytics**

* API endpoint usage tracking
* Function execution metrics
* Cache hit rates

**Cost Management**

**Development Phase**

* Free tier limitations understanding
* Development team preview deployments
* Minimal production infrastructure

**Production Phase**

* Pay-as-you-go scaling
* Function execution optimization
* Cache strategy optimization

**Security Implementation**

**Authentication**

* Implement authentication middleware
* API route protection
* Rate limiting implementation

**Data Protection**

* Environment variables management
* Secure backend communication
* Data encryption in transit

**Project Scope and Limitations - Detailed Analysis**

**Core Project Scope**

**Content Coverage**

1. **News Sources**

* Scope:
  + Top 20 English-language news portals globally
  + Major mainstream media outlets
  + Established digital news platforms
  + Official press releases
* Limitations:
  + No paywall-protected content
  + No subscription-based content
  + No regional/local news sources
  + No non-English content

1. **Content Categories**

* Scope:
  + Global politics and international relations
  + Business and economics
  + Technology and innovation
  + Sports (major events)
  + Entertainment (major news)
  + Science and research
  + Environmental news
  + Health and medicine
* Limitations:
  + No niche category coverage
  + Limited depth in specialized topics
  + No opinion pieces or editorials
  + No lifestyle content

**Functional Scope**

1. **Data Processing**

* Scope:
  + Real-time news aggregation
  + Cross-source fact verification
  + Entity extraction and relationship mapping
  + Sentiment analysis of public reactions
* Limitations:
  + Processing delay up to 5 minutes
  + Limited historical data analysis
  + No real-time video processing
  + Limited multimedia content analysis

1. **User Features**

* Scope:
  + TL;DR summaries
  + Interactive knowledge graphs
  + Social media sentiment analysis
  + Related video content
  + Basic search functionality
* Limitations:
  + No personalization features
  + No user accounts/profiles
  + No content saving/bookmarking
  + No user-generated content

**Technical Limitations**

**Infrastructure Constraints**

1. **Development Phase**

* Processing Capacity:
  + Maximum 1000 articles per day
  + Up to 100 concurrent users
  + Limited video processing capability
  + Basic caching implementation

1. **Initial Production Phase**

* Resource Limits:
  + Single VM instance limitations
  + Shared database resources
  + Limited concurrent processing
  + Basic scalability options

**Performance Boundaries**

1. **Response Times**

* Target Metrics:
  + API response: < 200ms
  + Page load: < 2 seconds
  + Graph rendering: < 1 second
  + Search results: < 500ms
* Limitations:
  + Complex graph queries may take longer
  + Large dataset processing delays
  + Video analysis processing time
  + Real-time update delays

1. **Data Retention**

* Storage Policies:
  + Active news: 30 days
  + Processed data: 90 days
  + Social media data: 7 days
  + Video analysis: 14 days

**Scalability Considerations**

**Growth Limitations**

1. **Initial Phase**

* User Capacity:
  + Up to 1,000 daily active users
  + Maximum 100 concurrent sessions
  + Limited batch processing
  + Basic caching implementation

1. **Technical Boundaries**

* System Constraints:
  + Single region deployment
  + Limited failover capability
  + Basic backup systems
  + Minimal redundancy

**Future Expansion Possibilities**

**Potential Extensions**

1. **Content Expansion**

* Additional Languages
* More news sources
* Specialized categories
* Historical data analysis

1. **Feature Enhancement**

* User accounts and personalization
* Advanced search capabilities
* Mobile applications
* API access for developers

**Not in Scope**

1. **Excluded Features**

* Content creation tools
* Social networking features
* Monetization systems
* Content management system

1. **Technical Exclusions**

* Multi-region deployment
* Full disaster recovery
* Enterprise-grade security
* Custom API development

**Compliance and Legal**

**Regulatory Scope**

1. **Data Handling**

* Public domain content only
* Basic GDPR compliance
* Standard security measures
* Public API usage only

1. **Content Rights**

* Fair use principles
* Attribution requirements
* Public content only
* No premium content