Alibaba Cloud DataWorks

プロダクト紹介

Document Version20190527

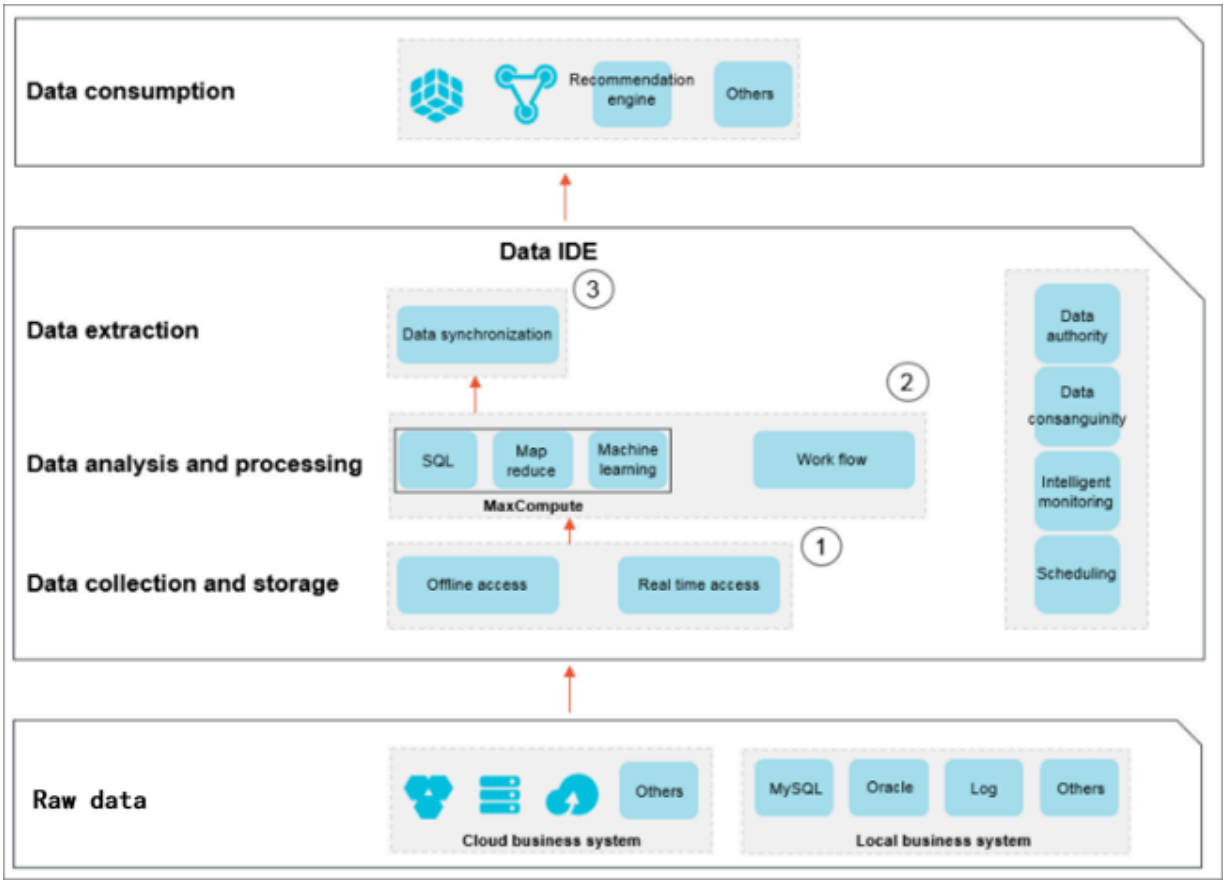
目次

1 DataWorks の概要

DataWorks は、PaaS (Platform as a Service) 製品として、Alibaba Cloud 重要なプラットフォームです。 DataWorks は、完全クラウド型のワークフロー機能と、ワンストップ型のデータ開発および管理インターフェイスを提供し、顧客企業のデータマイニングとデータ探索を支援します。

DataWorks は、MaxCompute をコアコンピューティングおよびストレージエンジンとして使用して、オフライン環境での強力なデータ処理、分析、マイニング能力を実現しています。 詳細は、「*MaxCompute の概要*」をご参照ください。

DataWorks はデータの転送と変換を容易にします。他のデータストレージサービスからデータをインポートして変換し、最終的に他のデータシステムへ伝送するためにデータを抽出することも可能です。 DataWorks におけるデータ分析の全体的な概要については、次の図をご参照ください。



機能

- 完全クラウド型スケジューリング機能

  DataWorks はパワフルなスケジュール機能を提供します。 DAG (Directed Acyclic Graph) のリレーションシップに基づき、時間や依存関係をベースとしたトリガーを設定し、毎日最大限の正確性をもって膨大なタスクを実行できます。 スケジュールの頻度は分単位、時間単位、日単位、週単位、月単位でサポートされています。

  完全クラウド型であるため、スケジュールに消費するサーバーのリソースを心配する必要がありません。 システムはテナントごとに分かれており、タスクを独立して実行します。
- 多様なタスクをサポート

  DataWorks はデータ同期、シェル、MaxCompute SQL 、MaxCompute MR など、多様なタスクをサポートします。 タスク間の依存関係により、複雑なデータ分析処理が実行できます。

  - MaxCompute の搭載により、DataWorks はパワフルなデータ変換機能を実現しており、高性能なビッグデータ分析処理を保証します。
  - データ同期について、DataWorks は強力なデータ統合機能を備え、20 種類以上ものデータソースをサポートし、安定した高効率なデータ転送を提供します。 詳細は、*Data Integration の概要*をご参照ください。
- ビジュアル開発

  本製品はビジュアルなコード開発とワークフローデザイン画面を提供します。 開発ツールの追加をしなくても、部品をドラッグ&ドロップして複雑なデータ分析タスクを構築できます。 インターネット接続とブラウザーがあれば、どこでも開発作業が可能です。
- 監視とアラーム

  運用センターで、タスクの監視と管理をビジュアルに行うことができます。運用センターは、タスク実行時に DAG フォーマットで全体状況を表示します。

  タスク障害の迅速な解決に向けたSMS アラーム通知を簡単に設定できます。

制約と制限

- DataWorks は Chrome 54 以降のバージョンのみサポートしています。
- 現在、 DataWorks は Alibaba Cloud のMaxCompute での SQL 操作のみをサポートしています。

2 基本概念

業務フロー

本ドキュメントでは、DataWorks の業務フロー、ソリューション、コンポーネント、タスク、インスタンス、送信、スクリプト、開発、リソース、機能、および出力名などの基本概念について説明します。

利点:

- 業務観点でデータコードの体系化を支援します。タスクの種類、多層的なサブディレクトリ (Alibaba Cloud の推奨は 4 層以下) によってコードを体系化します 。
- 最適化に向けた業務観点でのワークフロー概要の提供。
- 効率的な開発のための業務フローダッシュボードの提供。
- 業務フローをもとにしたリリース、メンテナンスの体系化。

ソリューション

DataWorks はカスタマイズ可能な業務フローソリューションを提供します。

利点:

- 多様な業務フローが利用可能になります。
- さまざまなソリューションで再利用可能な業務フローを提供できます。
- 没入型開発向けの包括的なソリューションを提供できです。

コンポーネント

コンポーネントは、複数のインプットおよびアウトプットパラメーターを持つ SQL コードテンプレートです。SQL コードは一般的に、1つまたはそれ以上のソースデータテーブルをフィルタリング、接続、集計、その他操作を行うことで処理されます。新規業務ニーズに応じて対象テーブルを処理します。 SQL の通常ロジックでは、コンポーネントを抽出し、コードの再利用機能を高めることが可能です。

タスク

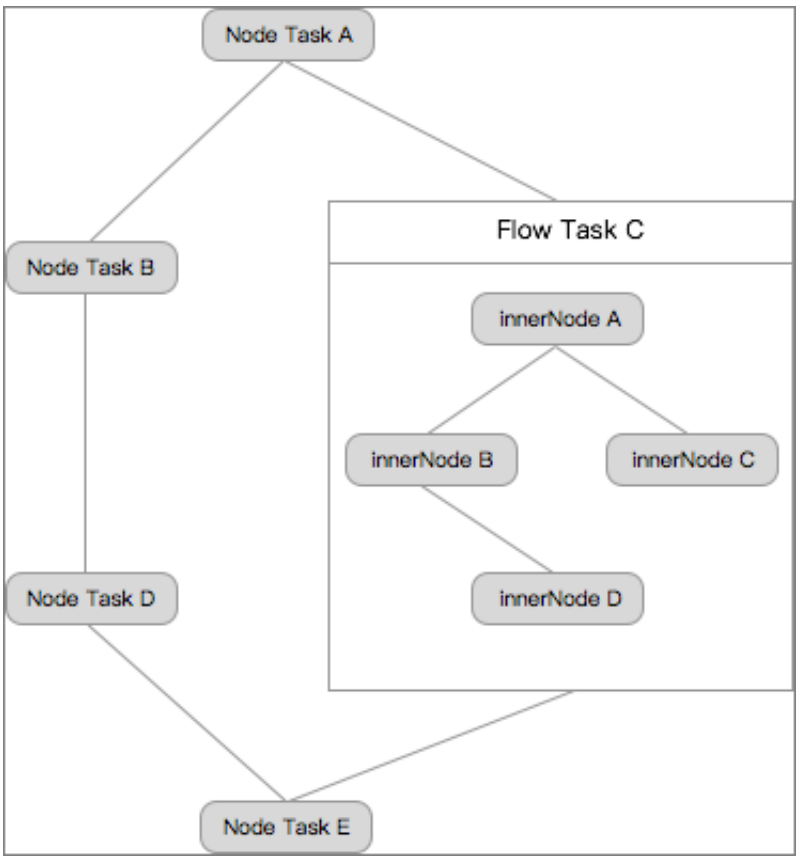
タスクは、さまざまなデータに対して実行される操作です。 多様なタスクアプリケーションに関する説明は以下のとおりです:

- データ同期ノードタスクを使用して、RDS からMaxCompute にデータをコピーします。
- データ変換のための MaxCompute SQL の実行にあたり、MaxCompute SQL ノードタスクを使用します。

・ フロータスクを使用して、いくつかの内部SQLノード間で一連のデータ変換を実行します。

各タスクにおいて、0 個以上のデータテーブル (データセット) が入力され、1 つ以上のデータテーブル (データセット) が出力されます。

タスクは、ノードタスク、フロータスク、および内部ノードに分かれます。 これらタスクの関係性については、次の図をご参照ください:



・ ノードタスクはデータに対して実行される操作です。 DAG (有向非循環グラフ) を形成するために、他のノードタスクとフロータスクに依存するように設定できます。
・ フロータスクは、ワークフロータスクを処理する内部ノードグループによって形成されるタスクです。 10 個未満のフロータスクの使用を推奨します。 フロータスクの内部ノードは、他のフロータスクやノードタスクに依存できません。 フロータスクは、他のフローやノードタスクに依存してDAGを形成するように設定できます。
・ 内部ノードはフロータスク内のノードです。 基本的に内部ノードはノードタスクと同じ機能を提供します。 内部ノードのスケジューリングサイクルは、フロータスクのスケジュール頻度を継承するため、独立した設定はできません。 ドラッグするだけで依存関係の設定が可能です。

データ操作は操作の種類から選択可能です。詳細はノードタイプの概要をご参照ください。

タスクスケジューリングのパラメーター設定については、スケジューリング設定をご参照ください。

インスタンス

タスクがシステムによってスケジュールされるか、または手動で設定されると、インスタンスが生成されます。 インスタンスは、特定の時刻にタスクによって実行されるスナップショットです。 インスタンスには、タスクの実行時間、実行状況、実行ログ、およびその他情報が含まれています。 例:

タスク 1 は、毎日 2:00 に実行されるように構成されているものとします。 この場合、スケジューリングシステムは、定期ノード タスクによって事前定義された時刻 23:30 に、毎日スナップショットを自動生成します。 つまり、タスク 1 のインスタンスは翌日の午前 2 時に実行されます。 アップストリームタスクが完了したことを検出すると、システムは翌日の午前 2 時にタスク 1 のインスタンスを自動的に実行します。

 注 :

O&M Center > Task O&M ページでタスクインスタンス情報を照会できます。

送信

送信とは、ノードタスク開発プロセスであり、スケジューリングシステムに公開される開発環境からのワークフロータスクです。 タスクが送信されると、そのコードとスケジューリング設定がスケジューリングシステムに同期され、スケジューリングシステムはその設定に従ってタスクをスケジュールします。

 注 :

送信されていないノードタスクとフロータスクは、スケジューリングシステムに入力しないでください。

スクリプト

スクリプトとは、データ分析のために提供されるコード記憶領域です。 スクリプトコードをスケジューリングシステムにリリースすることはできず、スケジューリングパラメーターを設定することもできません。 スクリプトコードは、データクエリとデータ分析にのみ使用できます。

リソースと関数

リソースと関数はどちらも MaxCompute の概念です。 詳細は「*MaxCompute リソース*」と「*MaxCompute 関数*」をご参照ください。

DataWorks では、リソースと機能の管理にインタフェースを使用しています。 他の MaxCompute メソッドで管理されているリソースや関数を DataWorks で照会することはできません。

出力名

出力名は各タスクにおける出力ポイントの名称です。 Alibaba Cloud アカウントのシングルテナント内で依存性を設定する場合、アップストリームタスクとダウンスリームタスクに接続する仮想エンティティです。

他のタスクに依存するアップストリームとダウンストリームが構成される設定となっている場合、その際の設定は出力名をベースにしなければなりません。 また、その際のタスクの出力名は、ダウンストリームノードのインプット名にする必要があります。

 注 :

タスクIDと類似するアウトプット名は、同一テナント内のタスクとは異なる固有の概念対象となります。 カスタム出力名の追加も可能ですが、同じテナント内では固有の出力ノード名にする必要があります。

3 シナリオ

インターネット ビッグデータ アプリケーションサービス向けクラウドプラットフォームの構築

特徴:

- 企業による基幹業務へのフォーカスを支援

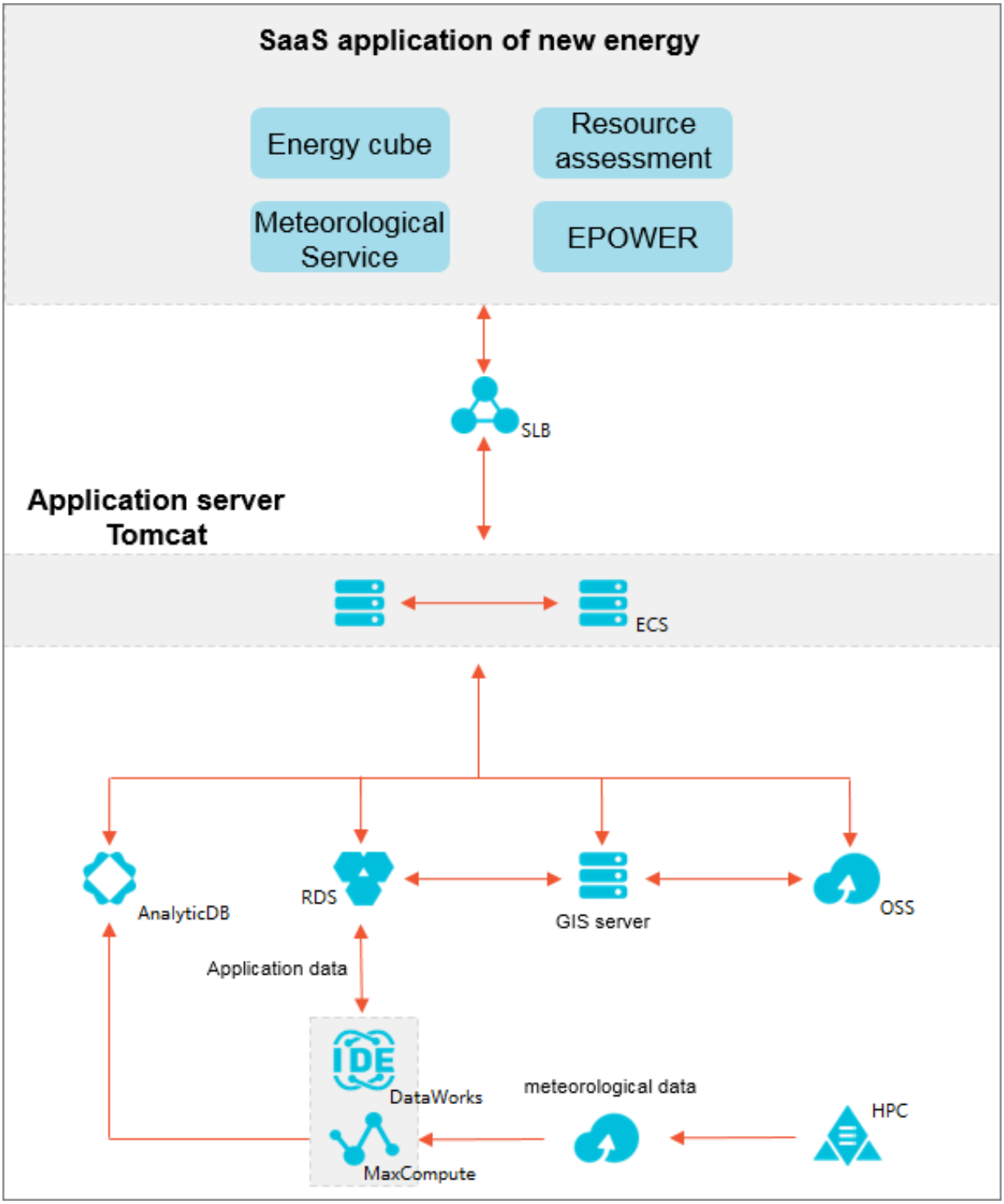
  すべての業務基盤を短時間で Alibaba Cloud へ移管し、膨大なリソースとともに業務の生産性を最適化します。 Alibaba Cloud の成熟した企業向けスケーリングソリューションによって、シームレスにスケーリングすることや関連事項への集中対応は不要となります。
- 投資費と運用保守費を削減

  オンプレミスのビッグデータ基盤に必要な物的リソース、労力、研究開発投資を大幅に削減できます。
- セキュリティと安定性

  クラウドへの完全データ移行は、DataWorks の包括的なサービス機能と安定した安全なパフォーマンスによって保証されています。

推奨する組み合わせ

DataWorks + AnalyticDB + MaxCompute



気象データクエリと広告事業ログ分析

特徴:

- **作業効率を向上**

  すべてのログデータは SQL 文で解析され、業務効率は 5 倍以上向上します。

- ストレージ利用率を改善

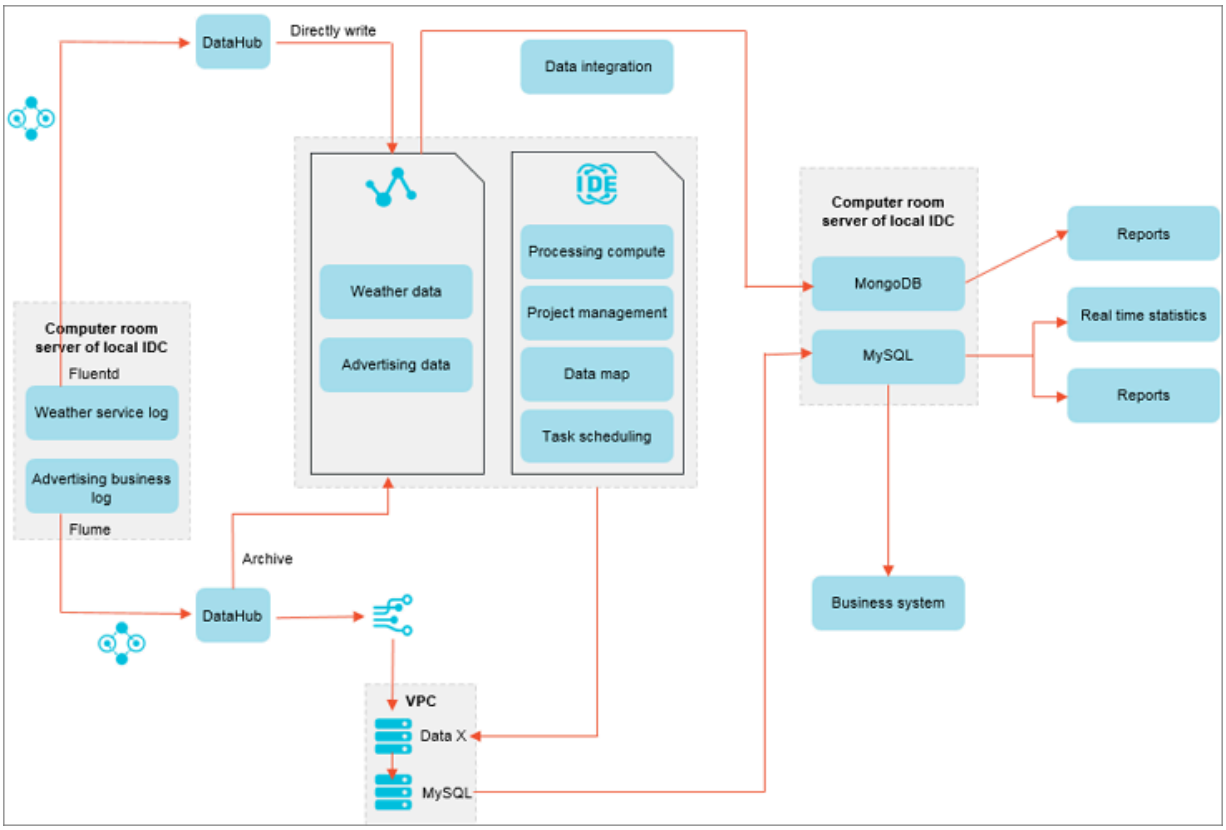
  DataWorks によってすべてのストレージとコンピューティング費用は 70% 削減し、パフォーマンスと安定性も向上します。

- ビッグデータ製品の使いやすさを向上

  MaxCompute は多様なオープンソースソフトウェアに対応するプラグインを提供しているため、クラウドへの容易なデータ移管が可能です。

推奨する組み合わせ

DataWorks + Data Integration + AnalyticDB + Quick BI + MaxCompute



細かな管理操作

- ビジネスインサイトの向上

  MaxCompute のコンピューティング機能により、数百万ものユーザーが細かな管理を行えるようになります。

- データ駆動型ビジネス

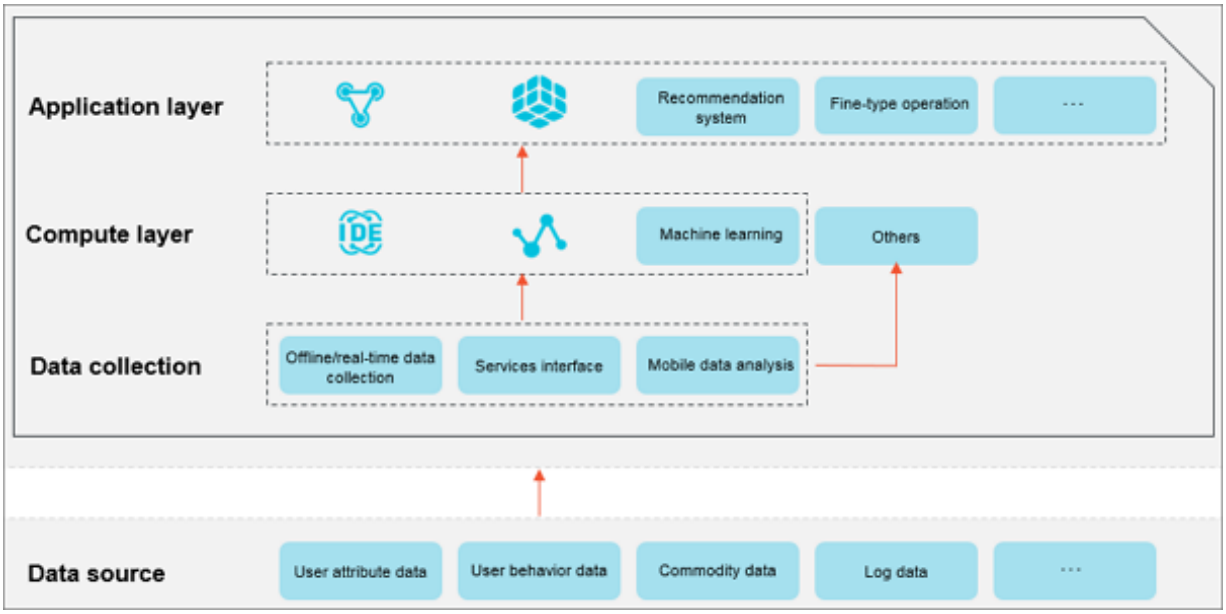
  DataWorks は、強化されたデータ分析機能と効果的な監視機能を提供することによってビジネスを強化します。

・ 業務要求に対する迅速な対応

DTplus エコシステムは新たな業務データの分析要求に迅速に対応します。

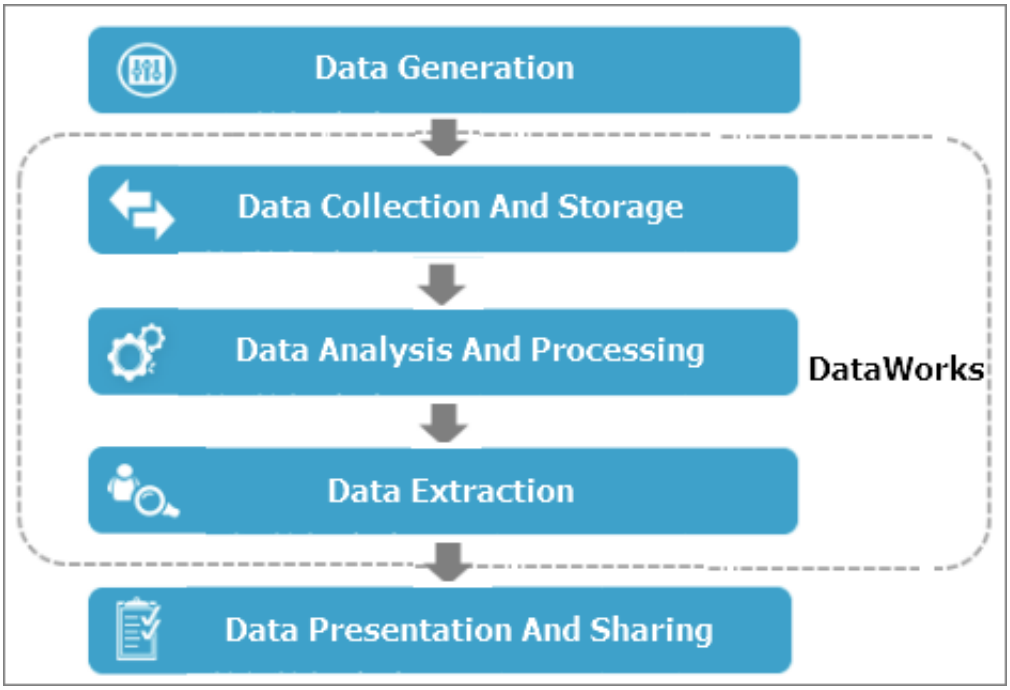
推奨する組み合わせ

DataWorks + Data integration + Quick BI + MaxCompute



4 データ開発プロセス

データ開発プロセスは、データ生成、データ収集と格納、データ分析と処理、データ抽出、データ表示と共有で構成されています。 プロセスについては、以下の図をご参照ください。



注 :

上の図では、点線枠内のデータ開発処理が Alibaba Cloud Big Data Platform で完了しています。

データ開発処理は以下のとおりです。

- データ生成

  業務システムは、毎日大量の構造化データを生成します。 データは、MySQL、Oracle、および RDS などの業務システムデータベースに格納されます。

- データ収集と格納

  データ分析のために MaxCompute の大容量データストレージと処理機能を用いるには、異なる業務システムのデータを MaxCompute に同期させる必要があります。

  DataWorks は、事前定義されたスケジューリング期間に沿って、業務システムのさまざまなデータを MaxCompute に同期するデータ統合サービスを提供します。

- **データ分析と処理**

  MaxCompute 上で処理 (MaxCompute_SQL やOPEN_MR) 、分析、マイニング (データ分析やデータマイニング) することで、有益な情報を見つけ出せます。

- **データ抽出**

  分析と処理が完了したデータは、以後の使用のためにビジネスシステムと同期させる必要があります。

- **データ表示と共有**

  最終的に、ビッグデータの分析と処理の結果がレポートや地理情報システム、その他アクセス可能な形式で表示され、共有されます。

5 簡易モードと標準モード

新しいバージョンの DataWorks では、簡易モードと標準モードの両方が導入されています。本ページでは、簡易モードと標準モードの違いについて説明します。

簡易モード

簡易モードとは、MaxCompute プロジェクトに対応し、開発および本番環境を設定できない DataWorks プロジェクトのことです。データ開発プロセスとテーブル権限を厳格に制御しないと簡易データ開発しか実行できません。

簡易モードの利点は、繰り返しが速く、コードが公開されずに送信できることです。これは効果を発揮します。

簡易モードのリスクは、開発ロールの特権が高すぎてこのプロジェクト配下のテーブルを削除できないことです。テーブル権限のリスクがあります。

標準モード

標準モードとは、2 つの MaxCompute プロジェクトに対応する DataWorks プロジェクトのことで、デュアル環境の開発と本番、コード開発仕様の改善、およびテーブル権限の厳密な制御を行うために立ち上げられます。 本番環境でのテーブル操作は禁止されており、本番テーブルのデータセキュリティは保証されています。

- すべてのタスク編集は開発環境でのみ実行でき、本番環境コードを直接変更することはできません。本番環境コードの安定性を確保するために、できる限り本番環境コードの変更エントリーを減らしてください。
- 開発環境は、タスクスケジューリングをデフォルトではオンにしません。環境プロジェクトのサイクル運用と資源を占有するための環境プロジェクトの本番の開発を避けることで、本番環境タスクの運用安定性はより保証されます。
- 本番環境はデフォルトの本番アカウントで実行されます。本番アカウントによって作成されたすべてのテーブルはメインアカウントに属します。開発プロセス中は本番テーブルを使用する必要があります。テーブルのすべては個別に適用される必要があり、テーブル権限の制御が向上します。

プロジェクトを作成するときは、標準モードとして [project mode] を選択し、プロジェクト名とプロジェクトの説明を入力します。残りの設定項目ではデフォルト値を選択します。

 注 :

本番環境の MaxCompute アクセス ID を個人アカウントに変更することはできません。そうでなければ、本番環境のデータセキュリティは保証されません。

6 バージョン履歴

DataWorks V2.0 のリリース

リリースバージョン: DataWorks V2.0

- リリース日: 2018 年 7 月 25 日
- リリース範囲: 中国 (上海) デプロイメントのみ
- リリース: DataWorks V2.0 をもとに DataWorks V2.0 は業務プロセスとコンポーネントを追加します。データ研究開発システムも改良し、開発環境と本番環境分離してデュアルプロジェクト開発をサポートでき、データ開発の仕様を保証し、エラーコードの発生を減らします。
- DataWorks V2.0 に関するより詳細な情報は「*DataWorks V2.0 FAQ と難度分析*」をご参照ください。

Dataworks 2.0 サポートリージョン

Dataworks 2.0 サポートリージョン

- 中国 (杭州)
- 中国 (上海)
- 中国 (北京)
- 中国 (深セン)

DataWorks V2.0 更新リスト

DataWorks V2.0 更新リスト					
DataWorks V2.0 はビジュアルインタラクションと データ開発モジュールの使用経験を向上しました。 更に、DataWorks V2.0 は高度な監視、データ保護、データ品質、データサービスを含む 4 つの新たなモジュールを提供しています。 最新バージョンへの円滑な移行を図るために、以下 DataWorks V2.0 の更新リストを記載します。					
モジュール名	サブモジュール	比較項目	DataWorks V1. 0	DataWorks V2.0	改善効果

DataWorks V2.0 更新リスト					
MaxCompute プロジェクト	プロジェクト管理モード	管理方法	1 つの DataWorks プロジェクトは 1 つの MaxCompute プロジェクトに対応しています。	DataWorks 標準モードの概念について。 標準モードにて、1 つの DataWorks プロジェクトは開発環境と本番環境の 2 つの MaxCompute プロジェクトに対応します。『シンプルモードと標準モード』をご参照ください。	リスクを隔離し、本番環境コードの安定性を確保します。
データ開発	タスク開発	全体の機能	シングルタスクの実行、ワークフローコードの書き込み、サイクルスケジューリングの設定。 完了後、オペレーションセンターに送信し、自動スケジューリングを実行。	· 名称変更: データ開発 · 新規: ソリューション、業務プロセスコンセプト · 削除: ワークフロー(コンセプト) · 最適化: より高度な SQL 編集、タスクサイクル設定、よりオープンな依存関係の設定。	1. SQL 編集: ユーザーフレンドリーで強力な SQL開発。 2. タスク管理: 業務プロセスとソリューションによって複雑な開発タスクを容易に管理できます。 3. タスクスケジューリング: オープンなスケジューリングシステムによって複雑な業務シナリオ処理が容易に。 4. その他機能: ユーザーの抱える問題を解決するための最適化された新しい機能。
		SQL の研究開発	単一のタスクまたはワークフローの形式でページに SQL コードを書き、それをテスト実行します。	コードハイライト、書式設定、高度な補完、エラー情報、テーブル構造表示、その他ユーザーフレンドリー機能などより高度な SQL エディタを提供します。 同時に、SQL の内部 [構造] をグラフ形式で可視化します。	
		ノード設定	シングルノードとワークフローモードを介した業務コードの統合	ワークフローの業務プロセスコンセプトについて 業務プロセスへのタスク統合が可能となり、要求に応じた業務プロセスの異なるリソース管理が可能 (全タスク、テーブル、リソース、機能は業務プロセスに属する必要あり) 。強い関連性で業務プロセス管理を統合し、 ソリューションでの業務プロセスを1ステップに統合可能。	

DataWorks V2.0 更新リスト				
		サイクル設定	ワークフローの全体的なサイクル設定は定期的な個人タスク設定に影響。	すべてのノードは単独設定が可能。アップストリームノードとダウンストリームノードによるスケジューリングサイクルの種類への影響なし。
		依存属性	ワークフロー間の依存関係に制限があります。	異なる業務プロセスでのタスクノードは依存関係になることができます。業務プロセスとの依存関係にする必要はありません。
	スクリプト開発	全体機能	定期的なタスク供給は、一時的な不定期データ処理に使用します。	同じ機能の場合、マニュアル業務プロセスと名前を変更します。
	リソース管理	全体機能	MaxCompute プロジェクト内で別のタブとしてすべてのリソース (jar、ファイル、アーカイブ等) を管理します。	ユーザーは管理用の多階層フォルダーを作成しながら、業務プロセスに関係するリソースに、業務プロセスのサブラベルとしてオンデマンドに加入することが可能。
	機能管理	全体機能	MaxCompute SQL 編集に必要なシステムとカスタム機能を管理する個別のタグ。	個別のタグまたはサブタブとして必要な機能を管理するワークプロセスでの全機能を管理します。
	テーブル照会	全体機能	コンテンツのプレビュー、参照、表示機能を備えた MaxCompute プロジェクト下ですべてのテーブルを表示します。	同じ

DataWorks V2.0 更新リスト			
テーブル管理 (new)	全体機能	なし	開発者向けに、テーブル管理、ライフサイクル設定、テーブル管理を利用します。カテゴリー、説明、フィールド、パーティション、非表示、表示、テーブル削除、などのテーブル管理機能をサポートします。
一時クエリ (new)	全体機能	なし	コードが要件に一致するかどうかのテストに使用されます。 送信、公開、スケジュール設定、およびパラメーター機能は含まれていません。
コンポーネント管理 (new)	全体機能	なし	多数の類似する再利用可能な SQL コードを、 SQL コードブロックやノードタスクで抽出。また、入力引数と出力引数の設定が可能で、その設定を多様な実務に適用可能です。
実行履歴 (new)	全体機能	なし	ローカルでの過去 3 日間分のタスク実行履歴を表示できます。タスク実行結果の閲覧と簡易なフィルタリングも可能です。
結果のフィルター (new)	全体機能	なし	フィルタリングによって期待される結果の取得が可能。Excel コンポーネントに統合する SQL の 結果を提供し、結果のプリント後、並び替えを行います。
リサイクルビン (new)	全体機能	なし	ユーザータスクの誤削除による業務上の損失を回避して、リサイクルビンの現行項目で削除されたノードを確認し、回復機能を提供できます。

DataWorks V2.0 更新リスト					
	グローバルコードの検索	全体機能	なし	不完全な文字列で MaxCompute SQL 、シェル、データ同期タスクの検索が可能で、閲覧または操作に必要なタスクを迅速に見つけることができます。	
	リリース機能	全体機能	DataWorksV1 の標準モードプロジェクトでの公開機能を保持。	名称の変更: プロジェクトクローン シンプルスキーマプロジェクトは、自動的に他のプロジェクトへタスクをクローンする機能を維持しています。	
運用保守センター	タスクリスト	機能	ノードタイプ、名前、および所有者に基づいてタスクを検索します。	業務プロセス、ソリューション、基準名を用いたタスク検索機能を追加します。	タスク運用は、業務観点でタスク開発インターフェース上の最新機能に一致。
	運用タスク	機能	ノードタイプ、名前、オーナー、業務日、実行日に基づいてタスクを検索します。	業務プロセス、ソリューション、基準名を用いたタスク検索機能を追加します。	
	警告	機能	監視アラームは、エラー、完了、不完全なイベントなどに基づいています。	基準の監視、アラームのインシデント、カスタムアラームの 3 つの機能を統合し、より高度で完全なアラームシステムを構築します。	

DataWorks V2.0 更新リスト		
高度な監視 (new)	「アラーム」は実行中のDataWorks タスクを監視、分析するシステムです。 高度な監視は、監視ルールやタスク操作にもとづき、警告の必要性や、いつ、誰に、どうやって警告を出すかを判断します。 また、自動的に最適な警告時間、警告方法、警告対象を選択します。	ワンストップアクセスで、クラウド上のデータ開発、データガバナンス (セキュリティ) を実現し、閉ループデータ共有の経験を提供します。
データ品質 (new)	「データ品質」は、さまざまな異種データソースの品質検証、通知、および管理サービスをサポートするワンストッププラットフォームです。 DQC は監視対象としてデータセットを使用し、MaxCompute データテーブルの監視、 DataHubのリアルタイムデータストリームを監視できます。 MaxCompute データが変更された場合、DQC はデータを確認して製品リンクをブロックし、データ汚染の拡散を防止します。 そのうえ、DQC は履歴結果を検証します。したがって、データ品質の分析、数値化が可能です。	
データサービス (new)	「データサービススタジオ」は迅速にデータAPI をデータテーブルから生成する機能を保有。ユーザーが既存API をデータサービスプラットフォームに登録可能になり、一元管理や公開を実現できます。 また、データサービスは API ゲートウェイに接続されています。 ワンクリックで、 API を API ゲートウェイにデプロイ可能です。 データサービスは API ゲートウェイと互換性があり、セキュアで安定した低コストの簡便なデータ共有サービスを提供します。	
データアンブレラ (new)	「データセキュリティ保護 」は特定のデータ資産検証、機密データの検出、データの分類、モニターアクセス機能、識別、警告、および監査を提供します。	